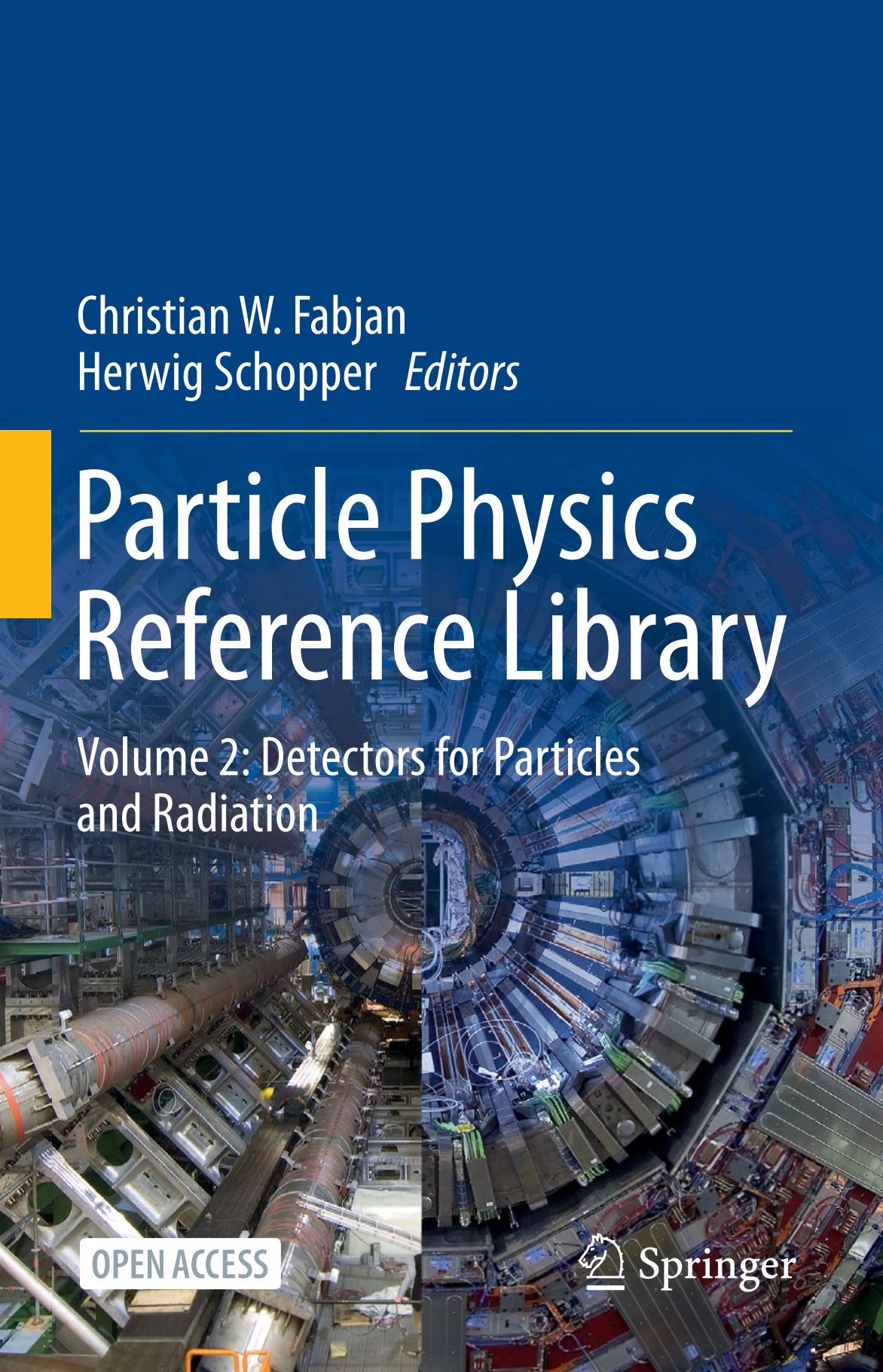
Christian W. Fabjan
Herwig Schopper *Editors*

Particle Physics Reference Library

Volume 2: Detectors for Particles and Radiation



OPEN ACCESS

Springer

Particle Physics Reference Library

Christian W. Fabjan • Herwig Schopper
Editors

Particle Physics Reference Library

Volume 2: Detectors for Particles and Radiation

Springer Open

Editors
Christian W. Fabjan
Austrian Academy of Sciences and
University of Technology
Vienna, Austria

Herwig Schopper
CERN
Geneva, Switzerland

ISBN 978-3-030-35317-9 ISBN 978-3-030-35318-6 (eBook)
https://doi.org/10.1007/978-3-030-35318-6

This book is an open access publication.

© The Editor(s) (if applicable) and The Author(s) 2011, 2020
Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.
The images or other third party material in this book are included in the book's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

For many years, the *Landolt-Börnstein—Group I Elementary Particles, Nuclei and Atoms*, Vol. 21A (*Physics and Methods. Theory and Experiments*, 2008), Vol. 21B1 (*Elementary Particles. Detectors for Particles and Radiation*. Part 1: *Principles and Methods*, 2011), Vol. 21B2 (*Elementary Particles. Detectors for Particles and Radiation*. Part 2: *Systems and Applications*), and Vol. 21C (*Elementary Particles. Accelerators and Colliders*, 2013), has served as a major reference work in the field of high-energy physics.

When, not long after the publication of the last volume, open access (OA) became a reality for HEP journals in 2014, discussions between Springer and CERN intensified to find a solution for the "Labö" which would make the content available in the same spirit to readers worldwide. This was helped by the fact that many researchers in the field expressed similar views and their readiness to contribute.

Eventually, in 2016, on the initiative of Springer, CERN and the original Labö volume editors agreed in tackling the issue by proposing to the contributing authors a new OA edition of their work. From these discussions a compromise emerged along the following lines: transfer as much as possible of the original material into open access; add some new material reflecting new developments and important discoveries, such as the Higgs boson; and adapt to the conditions due to the change from copyright to a CC BY 4.0 license.

Some authors were no longer available for making such changes, having either retired or, in some cases, deceased. In most such cases, it was possible to find colleagues willing to take care of the necessary revisions. A few manuscripts could not be updated and are therefore not included in this edition.

We consider that this new edition essentially fulfills the main goal that motivated us in the first place—there are some gaps compared to the original edition, as explained, as there are some entirely new contributions. Many contributions have been only minimally revised in order to make the original status of the field available as historical testimony. Others are in the form of the original contribution being supplemented with a detailed appendix relating recent developments in the field. However, a substantial fraction of contributions has been thoroughly revisited by their authors resulting in true new editions of their original material.

We would like to express our appreciation and gratitude to the contributing authors, to the colleagues at CERN involved in the project, and to the publisher, who has helped making this very special endeavor possible.

Vienna, Austria — Christian W. Fabjan
Geneva, Switzerland — Stephen Myers
Geneva, Switzerland — Herwig Schopper
July 2020

Contents

1 Introduction 1
Christian W. Fabjan and Herwig Schopper

2 The Interaction of Radiation with Matter 5
Hans Bichsel and Heinrich Schindler

3 Scintillation Detectors for Charged Particles and Photons 45
P. Lecoq

4 Gaseous Detectors 91
H. J. Hilke and W. Riegler

5 Solid State Detectors 137
G. Lutz and R. Klanner

6 Calorimetry 201
C. W. Fabjan and D. Fournier

7 Particle Identification: Time-of-Flight, Cherenkov and Transition Radiation Detectors 281
Roger Forty and Olav Ullaland

8 Neutrino Detectors 337
Leslie Camilleri

9 Nuclear Emulsions 383
Akitaka Ariga, Tomoko Ariga, Giovanni De Lellis, Antonio Ereditato, and Kimio Niwa

10 Signal Processing for Particle Detectors 439
V. Radeka

11 Detector Simulation 485
J. Apostolakis

12 Triggering and High-Level Data Selection 533
W. H. Smith

13 Pattern Recognition and Reconstruction 555
R. Frühwirth, E. Brondolin, and A. Strandlie

14 Distributed Computing 613
Manuel Delfino

15 Statistical Issues in Particle Physics 645
Louis Lyons

16 Integration of Detectors into a Large Experiment: Examples from ATLAS and CMS 693
Daniel Froidevaux

17 Neutrino Detectors Under Water and Ice 785
Christian Spiering

18 Spaceborne Experiments 823
Roberto Battiston

19 Cryogenic Detectors 871
Klaus Pretzl

20 Detectors in Medicine and Biology 913
P. Lecoq

21 Solid State Detectors for High Radiation Environments 965
Gregor Kramberger

22 Future Developments of Detectors 1035
Ties Behnke, Karsten Buesser, and Andreas Mussgiller

About the Editors

Christian W. Fabjan is an experimental particle physicist, who spent the major part of his career at CERN, with leading involvement in several of the major CERN programs. At the Intersecting Storage Rings, he concentrated on strong interaction physics and the development of new experimental techniques and followed at the Super Synchrotron with experiments in the Relativistic Heavy Ion program. At the Large Hadron Collider, he focused on the development of several experimental techniques and participated in the ALICE experiment as Technical Coordinator. He is affiliated with the Vienna University of Technology and was, most recently, leading the institute of High Energy Physics of the Austrian Academy of Sciences.

Herwig Schopper joined as a research associate at CERN since 1966 and returned in 1970 as leader of the Nuclear Physics Division and went on to become a member of the directorate responsible for the coordination of CERN's experimental program. He was chairman of the ISR Committee at CERN from 1973 to 1976 and was elected as member of the Scientific Policy Committee in 1979. Following Léon Van Hove and John Adams' years as Director-General for research and executive Director-General, Schopper became the sole Director-General of CERN in 1981.

Schopper's years as CERN's Director-General saw the construction and installation of the Large Electron-Positron Collider (LEP) and the first tests of four detectors for the LEP experiments. Several facilities (including ISR, BEBC, and EHS) had to be closed to free up resources for LEP.

Chapter 1
Introduction

Christian W. Fabjan and Herwig Schopper

Enormous progress has been achieved during the last three decades in the understanding of the microcosm. This was possible by a close interplay between new theoretical ideas and precise experimental data. The present state of our knowledge has been summarised in Volume I/21A "Theory and Experiments". This Volume I/21B is devoted to detection methods and techniques and data acquisition and handling.

The rapid increase of our knowledge of the microcosm was possible only because of an astonishingly fast evolution of detectors for particles and photons. Since the early days of scintillation screens and Geiger counters a series of completely new detector concepts was developed. They are based on imaginative ideas, sometimes even earning a Nobel Prize, combined with sophisticated technological developments. It might seem surprising that the exploration of an utterly abstract domain like particle physics, requires the most advanced techniques, but this makes the whole field so attractive.

The development of detectors was above all pushed by the requirements of particle physics. In order to explore smaller structures one has to use finer probes, i.e. shorter wavelengths implying higher particle energies. This requires detectors for high-energy particles and photons. At the same time one has to cope with the quantum-mechanical principle that cross sections for particle interactions have a tendency to fall with increasing interaction energy. Therefore accelerators or colliders have to deliver not only higher energies but at the same time also higher collision rates. This implies that detectors must sustain higher rates. This problem is aggravated by the fact that the high-energy frontier is at present linked to hadron

C. W. Fabjan (✉)
Austrian Academy of Sciences and University of Technology, Vienna, Austria
e-mail: Chris.Fabjan@cern.ch

H. Schopper
CERN, Geneva, Switzerland

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_1

collisions. Electron-positron colliders are characterised by events with relatively few outgoing particles since two pointlike particles collide and the strong interaction is negligible in such reactions. After the shutdown of LEP in 2000 the next electron-positron collider is far in the future and progress is now depending on proton-proton collisions at the LHC at CERN or heavy ion colliders, e.g. GSI, Germany, RHIC at BNL in the USA and also LHC. Protons are composite particles containing quarks and gluons and hence proton collisions produce very complicated events with many hundreds of particles. Consequently, detectors had to be developed which are able to cope with extremely high data rates and have to resist high levels of irradiation. Such developments were in particular motivated by the needs of the LHC experiments.

It seems plausible that accelerators and colliders have to grow in size with increasing energy. But why have detectors to be so large? Their task is to determine the direction of emitted particles, measure their momenta or energy and in some cases their velocity which together with the momentum allows to determine their mass and hence to identify the nature of the particle.

The most precise method to measure the momentum of charged particles is to determine their deflection in a magnetic field which is proportional to $B \cdot l$ where B is the magnetic field strength and l the length of the trajectory in the magnetic field. Of course, it is also determined by the spatial resolution of the detector to determine the track. To attain the highest possible precision superconducting coils are used in most experiments to produce a large B. Great efforts have been made to construct detectors with a spatial resolution down to the order of several microns. But even then track lengths l of the order of several meters are needed to measure momenta with a precision of about 1% of particles with momenta of several 100 GeV/c. This is the main reason why experiments must have extensions of several meters and weigh thousands of tons.

Another possibility to determine the energy of particles are so-called "calorimeters". This name is misleading since calorimeters have nothing to do with calorific measurements but this name became ubiquitous to indicate that the total energy of a particle is measured. The measurement is done in the following way. A particle hits the material of the detector, interacts with an atom, produces secondary particles which, if sufficiently energetic, generate further particles, leading to a whole cascade of particles of ever decreasing energies. The energy deposited in the detector material can be measured in various ways. If the material of the detector is a scintillator (crystal, liquid or gas), the scintillating light is approximately proportional to the deposited energy and it can be observed by, e.g., photomultipliers. Alternatively, the ionisation produced by the particle cascade can be measured by electrical means.

In principle two kinds of calorimeters can be distinguished. Electrons and photons produce a so-called electromagnetic cascade due to electromagnetic interactions. Such cascades are relatively small both in length and in lateral dimension. Hence electromagnetic calorimeters can consist of a homogenous detector material containing the whole cascade. Incident hadrons, however, produce in the cascade also a large number of neutrons which can travel relatively long ways before losing their energy and therefore hadronic cascades have large geometrical extensions even

in the densest materials (of the order few meters in iron). Therefore the detectors for hadronic cascades are composed of a sandwich of absorber material interspersed with elements to detect the deposited energy. In such a device, only a certain fraction of the total energy is sampled. The challenge of the design consists in making this fraction as much as possible proportional to the total energy. The main advantage of calorimeters, apart from the sensitivity to both charged and neutral particles, is that their size increases only logarithmically with the energy of the incident particle, hence much less than for magnetic spectrometers, albeit with an energy resolution inferior to magnetic spectrometers below about 100 GeV. They require therefore comparatively little space which is of paramount importance for colliders where the solid angle around the interaction area has to be covered in most cases as fully as possible.

Other detectors have been developed for particular applications, e.g. for muon and neutrino detection or the observation of cosmic rays in the atmosphere or deep underground/water. Experiments in space pose completely new problems related to mechanical stability and restrictions on power consumption and consumables.

The main aim in the development of all these detectors is higher sensitivity, better precision and less influence by the environment. Obviously, reduction of cost has become a major issue in view of the millions of detector channels in most modern experiments.

New and more sophisticated detectors need better signal processing, data acquisition and networking. Experiments at large accelerators and colliders pose special problems dictated by the beam properties and restricted space. Imagination is the key to overcome such challenges.

Experiments at accelerators/colliders and for the observation of cosmic rays have become big projects involving hundreds or even thousands of scientists and the time from the initial proposal to data taking may cover one to two decades. Hence it is sometimes argued that they are not well adapted for the training of students. However, the development of a new detector is subdivided in a large number of smaller tasks (concept of the detector, building prototypes, testing, computer simulations and preparation of the data acquisition), each lasting only a few years and therefore rather well suited for a master or PhD thesis. The final "mass production" of many detection channels in the full detector assembly, however, is eventually transferred to industry. These kinds of activities may in some cases have little to do with particle physics itself, but they provide an excellent basis for later employment in industry. Apart from specific knowledge, e.g., in vacuum, magnets, gas discharges, electronics, computing and networking, students learn how to work in the environment of a large project respecting time schedules and budgetary restrictions—and perhaps even most important to be trained to work in an international environment.

Because the development of detectors does not require the resources of a large project but can be carried out in a small laboratory, most of these developments are done at universities. Indeed most of the progress in detector development is due to universities or national laboratories. However, when it comes to plan a large experiment these originally individual activities are combined and coordinated

which naturally leads to international cooperation between scientists from different countries, political traditions, creeds and mentalities. To learn how to adapt to such an international environment represents a human value which goes much beyond the scientific achievements.

The stunning success of the "Standard Model of particle physics" also exhibits with remarkable clarity its limitations. The many open fundamental issues—origin of CP-violation, neutrino mass, dark matter and dark energy, to name just few—are motivating a vast, multi-faceted research programme for accelerator- and non-accelerator based, earth- and space-based experimentation. This has led to a vigorous R&D in detectors and data handling.

This revised edition provides an update on these developments over the past 7–9 years.

We gratefully acknowledge the very constructive collaboration with the authors of the first edition, in several cases assisted by additional authors. May this Open Access publication reach a global readership, for the benefit of science.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 2
The Interaction of Radiation with Matter

Hans Bichsel and Heinrich Schindler

2.1 Introduction

The detection of charged particles is usually based on their electromagnetic interactions with the electrons and nuclei of a detector medium. Interaction with the Coulomb field of the nucleus leads to deflections of the particle trajectory (multiple scattering) and to radiative energy loss (bremsstrahlung). Since the latter, discussed in Sect. 2.4.1, is inversely proportional to the particle mass squared, it is most significant for electrons and positrons.

"Heavy" charged particles (in this context: particles with a mass M exceeding the electron mass m) passing through matter lose energy predominantly through collisions with electrons. Our theoretical understanding of this process, which has been summarised in a number of review articles [1–7] and textbooks [8–13], is based on the works of some of the most prominent physicists of the twentieth century, including Bohr [14, 15], Bethe [16, 17], Fermi [18, 19], and Landau [20].

After outlining the quantum-mechanical description of single collisions in terms of the double-differential cross section $\mathrm{d}^2\sigma/(\mathrm{d}E\mathrm{d}q)$, where E and q are the energy transfer and momentum transfer involved in the collision, Sect. 2.3 discusses algorithms for the quantitative evaluation of the single-differential cross section

The author Hans Bichsel is deceased at the time of publication.

H. Bichsel · H. Schindler (✉)
CERN, Geneva, Switzerland
e-mail: Heinrich.Schindler@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_2

$\mathrm{d}\sigma/\mathrm{d}E$ and its moments. The integral cross section (zeroth moment), multiplied by the atomic density N, corresponds to the charged particle's inverse mean free path λ^{-1} or, in other words, the average number of collisions per unit track length,

$$\lambda^{-1} = M_0 = N \int_{E_{\min}}^{E_{\max}} \frac{\mathrm{d}\sigma}{\mathrm{d}E} \mathrm{d}E. \tag{2.1}$$

The stopping power $\mathrm{d}E/\mathrm{d}x$, i.e. the average energy loss per unit track length, is given by the first moment,

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = M_1 = N \int_{E_{\min}}^{E_{\max}} E \frac{\mathrm{d}\sigma}{\mathrm{d}E} \mathrm{d}E. \tag{2.2}$$

The integration limits $E_{\min,\,\max}$ are determined by kinematics. Due to the stochastic nature of the interaction process, the number of collisions and the sum Δ of energy losses along a particle track are subject to fluctuations. Section 2.5 deals with methods for calculating the probability density distribution $f(\Delta, x)$ for different track lengths x. The energy transfer from the incident particle to the electrons of the medium typically results in excitation and ionisation of the target atoms. These observable effects are discussed in Sect. 2.6.

As a prologue to the discussion of charged-particle collisions, Sect. 2.2 briefly reviews the principal photon interaction mechanisms in the X-ray and gamma ray energy range.

Throughout this chapter, we attempt to write all expressions in a way independent of the system of units (cgs or SI), by using the fine structure constant $\alpha \sim 1/137$. Other physical constants used occasionally in this chapter include the Rydberg energy $\mathrm{Ry} = \alpha^2 mc^2/2 \sim 13.6\,\mathrm{eV}$, and the Bohr radius $a_0 = \hbar c/\left(\alpha mc^2\right) \sim 0.529\,\text{Å}$. Cross-sections are quoted in barn ($1\,\mathrm{b} = 10^{-24}\,\mathrm{cm}^2$).

2.2 Photon Interactions

Photons interact with matter via a range of mechanisms, which can be classified according to the type of target, and the effect of the interaction on the photon (absorption or scattering) [9, 21]. At energies beyond the ultraviolet range, the dominant processes are photoelectric absorption (Sect. 2.2.1), Compton scattering (Sect. 2.2.2), and pair production (Sect. 2.2.3). As illustrated in Fig. 2.1, photoabsorption constitutes the largest contribution to the total cross section at low photon energies, pair production is the most frequent interaction at high energies, and Compton scattering dominates in the intermediate energy range.

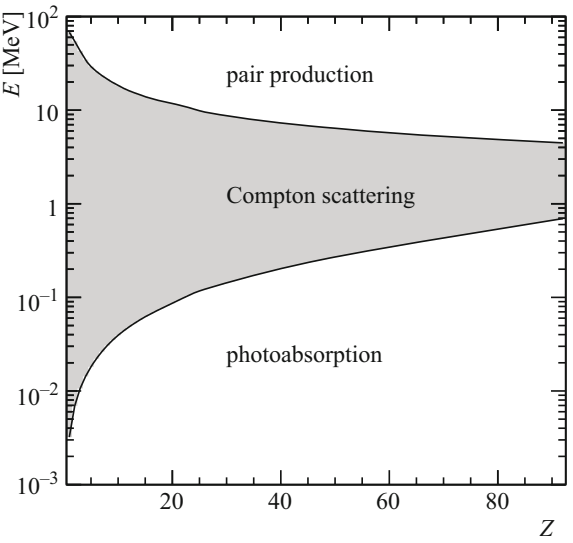


Fig. 2.1 The lower curve shows, as a function of the atomic number Z of the target material, the photon energy E below which photoelectric absorption is the most probable interaction mechanism, while the upper curve shows the energy above which pair production is the most important process. The shaded region between the two curves corresponds to the domain where Compton scattering dominates. The cross sections are taken from the NIST XCOM database [24]

Detailed descriptions of these processes can be found, for instance, in Refs. [8–10, 12, 22, 23]. The focus of this section is on photoabsorption, the description of which (as will be discussed in Sect. 2.3) is related to that of inelastic charged particle collisions in the regime of low momentum transfer.

2.2.1 Photoabsorption

In a photoelectric absorption interaction, the incident photon disappears and its energy is transferred to the target atom (or group of atoms). The intensity I of a monochromatic beam of photons with energy E thus decreases exponentially as a function of the penetration depth x in a material,

$$I(x) = I_0 \mathrm{e}^{-\mu x},$$

where the attenuation coefficient μ is proportional to the atomic density N of the medium and the photoabsorption cross section σ_γ,

$$\mu(E) = N\sigma_\gamma(E).$$

Let us first consider a (dipole-allowed) transition between the ground state $|0\rangle$ of an atom and a discrete excited state $|n\rangle$ with excitation energy E_n. The integral photoabsorption cross section of the line is given by

$$\int \sigma_\gamma^{(n)}(E)\,\mathrm{d}E = \frac{2\pi^2\alpha(\hbar c)^2}{mc^2} f_n.$$

The dimensionless quantity

$$f_n = \frac{2mc^2}{3\,(\hbar c)^2} E_n |\langle n| \sum_{j=1}^{Z} \mathbf{r}_i |0\rangle|^2, \tag{2.3}$$

with the sum extending over the electrons in the target atom, is known as the dipole oscillator strength (DOS). Similarly, transitions to the continuum are characterised by the dipole oscillator strength density $\mathrm{d}f/\mathrm{d}E$, and the photoionisation cross section $\sigma_\gamma\,(E)$ is given by

$$\sigma_\gamma\,(E) = \frac{2\pi^2\alpha\,(\hbar c)^2}{mc^2} \frac{\mathrm{d}f\,(E)}{\mathrm{d}E}. \tag{2.4}$$

The dipole oscillator strength satisfies the Thomas-Reiche-Kuhn (TRK) sum rule,

$$\sum_n f_n + \int \mathrm{d}E\, \frac{\mathrm{d}f\,(E)}{\mathrm{d}E} = Z. \tag{2.5}$$

For most gases, the contribution of excited states ($\sum f_n$) to the TRK sum rule is a few percent of the total, e.g. ~5% for argon and ~7% for methane [25, 26].

As can be seen from Fig. 2.2, the photoabsorption cross section reflects the atomic shell structure. Evaluated atomic and molecular photoabsorption cross

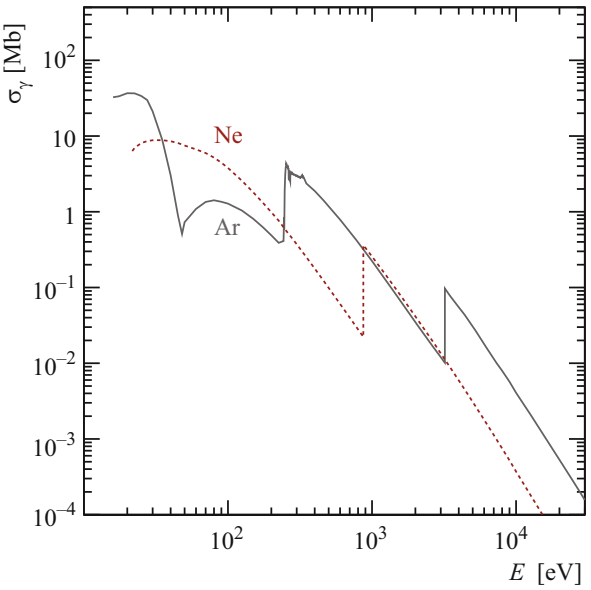


Fig. 2.2 Photoabsorption cross sections of argon (solid curve) and neon (dashed curve) as a function of the photon energy E [25, 26]

sections (both for discrete excitations as well as transitions to the continuum) for many commonly used gases are given in the book by Berkowitz [25, 26].

At energies sufficiently above the ionisation threshold, the molecular photoabsorption cross section is, to a good approximation, given by the sum of the photoabsorption cross sections of the constituent atoms. A comprehensive compilation of atomic photoabsorption data (in the energy range between ∼30 eV and 30 keV) can be found in Ref. [27]. Calculations for energies between 1 and 100 GeV are available in the NIST XCOM database [24]. Calculated photoionisation cross sections for individual shells can be found in Refs. [28–30]. At high energies, i.e. above the respective absorption edges, photons interact preferentially with inner-shell electrons. The subsequent relaxation processes (emission of fluorescence photons and Auger electrons) are discussed in Sect. 2.6.

The response of a solid with atomic number Z to an incident photon of energy $E = \hbar\omega$ is customarily described in terms of the complex dielectric function $\varepsilon(\omega) = \varepsilon_1(\omega) + \mathrm{i}\varepsilon_2(\omega)$. The oscillator strength density is related to $\varepsilon(\omega)$ by

$$\frac{\mathrm{d}f\,(E)}{\mathrm{d}E} = E\frac{2Z}{\pi\left(\hbar\Omega_p\right)^2}\frac{\varepsilon_2\,(E)}{\varepsilon_1^2\,(E)+\varepsilon_2^2\,(E)} = E\frac{2Z}{\pi\left(\hbar\Omega_p\right)^2}\mathrm{Im}\left(\frac{-1}{\varepsilon\,(E)}\right), \tag{2.6}$$

where

$$\hbar\Omega_p = \sqrt{\frac{4\pi\alpha\,(\hbar c)^3\,NZ}{mc^2}} \tag{2.7}$$

is the plasma energy of the material, which depends only on the electron density NZ. In terms of the dielectric loss function $\mathrm{Im}\,(-1/\varepsilon)$, the TRK sum rule reads

$$\int \mathrm{d}E\ \mathrm{Im}\left(\frac{-1}{\varepsilon\,(E)}\right)E = \frac{\pi}{2}\left(\hbar\Omega_p\right)^2. \tag{2.8}$$

Compilations of evaluated optical data for semiconductors are available in Ref. [32], and for solids in general in Ref. [31]. As an example, Fig. 2.3 shows the dielectric loss function of silicon, a prominent feature of which is the peak at ∼17 eV, corresponding to the plasma energy of the four valence (M-shell) electrons.

2.2.2 *Compton Scattering*

Compton scattering refers to the collision of a photon with a weakly bound electron, whereby the photon transfers part of its energy to the electron and is deflected with respect to its original direction of propagation. We assume in the following that the target electron is free and initially at rest, which is a good approximation if the photon energy E is large compared to the electron's binding energy. Due to

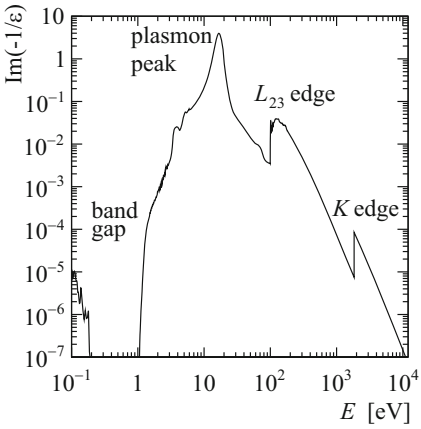


Fig. 2.3 Dielectric loss function $\mathrm{Im}\,(-1/\varepsilon\,(E))$ of solid silicon [31] as a function of the photon energy E

conservation of energy and momentum, the photon energy E' after the collision and the scattering angle θ of the photon are then related by

$$E' = \frac{mc^2}{1 - \cos\theta + (1/u)}, \tag{2.9}$$

where $u = E/\left(mc^2\right)$ is the photon energy (before the collision) in units of the electron rest energy.

The kinetic energy $T = E - E'$ imparted to the electron is largest for a head-on collision ($\theta = \pi$) and the energy spectrum of the recoil electrons consequently exhibits a cut-off (Compton edge) at

$$T_{\max} = E\frac{2u}{1+2u}.$$

The total cross section (per electron) for the Compton scattering of an unpolarised photon by a free electron at rest, derived by Klein and Nishina in 1929 [33], is given by

$$\sigma^{(\mathrm{KN})} = 2\pi\left(\frac{\alpha\hbar c}{mc^2}\right)^2\left(\frac{1+u}{u^2}\left[\frac{2\,(1+u)}{1+2u} - \frac{\ln\,(1+2u)}{u}\right] + \frac{\ln\,(1+2u)}{2u} - \frac{1+3u}{(1+2u)^2}\right). \tag{2.10}$$

At low energies ($u \ll 1$), the Klein-Nishina formula (2.10) is conveniently approximated by the expansion [34]

$$\sigma^{(\mathrm{KN})} = \underbrace{\frac{8\pi}{3}\left(\frac{\alpha\hbar c}{mc^2}\right)^2}_{\text{Thomson cross section}}\ \frac{1}{(1+2u)^2}\left(1 + 2u + \frac{6}{5}u^2 + \ldots\right),$$

while at high energies ($u \gg 1$) the approximation [8, 10, 22]

$$\sigma^{(\mathrm{KN})} \sim \pi \left(\frac{\alpha\hbar c}{mc^2}\right)^2 \frac{1}{u}\left(\ln(2u) + \frac{1}{2}\right)$$

can be used.

The angular distribution of the scattered photon is given by the differential cross section

$$\frac{\mathrm{d}\sigma^{(\mathrm{KN})}}{\mathrm{d}(\cos\theta)} = \pi \left(\frac{\alpha\hbar c}{mc^2}\right)^2 \left[\frac{1}{1+u(1-\cos\theta)}\right]^2 \left(\frac{1+\cos^2\theta}{2}\right) \times \left(1 + \frac{u^2(1-\cos\theta)^2}{(1+\cos^2\theta)[1+u(1-\cos\theta)]}\right),$$

which corresponds to a kinetic energy spectrum [22]

$$\frac{\mathrm{d}\sigma^{(\mathrm{KN})}}{\mathrm{d}T} = \pi \left(\frac{\alpha\hbar c}{mc^2}\right)^2 \frac{1}{u^2 mc^2}\left(2 + \left(\frac{T}{E-T}\right)^2 \left[\frac{1}{u^2} + \frac{E-T}{E} - \frac{2(E-T)}{uT}\right]\right)$$

of the target electron.

The cross section for Compton scattering off an atom scales roughly with the number of electrons in the atom and, assuming that the photon energy is large compared to the atomic binding energies, may be approximated by

$$\sigma^{(\mathrm{Compton})} \sim Z\sigma^{(\mathrm{KN})}.$$

Methods for including the effects of the binding energy and the internal motion of the orbital electrons in calculations of atomic Compton scattering cross sections are discussed, for instance, in Ref. [35].

2.2.3 *Pair Production*

For photon energies exceeding $2mc^2$, an interaction mechanism becomes possible where the incoming photon disappears and an electron-positron pair, with a total energy equal to the photon energy E, is created. Momentum conservation requires this process, which is closely related to bremsstrahlung (Sect. 2.4.1), to take place in the electric field of a nucleus or of the atomic electrons. In the latter case, kinematic constraints impose a threshold of $E > 4mc^2$.

At high photon energies, the electron-positron pair is emitted preferentially in the forward direction and the absorption coefficient due to pair production can be approximated by

$$\mu = N\sigma^{(\text{pair production})} = \frac{7}{9}\frac{1}{X_0},$$

where X_0 is a material-dependent parameter known as the radiation length (see Sect. 2.4.1). More accurate expressions are given in Ref. [8]. Tabulations of calculated pair-production cross sections can be found in Ref. [36] and are available online [24].

2.3 Interaction of Heavy Charged Particles with Matter

The main ingredient for computing the energy loss of an incident charged particle due to interactions with the electrons of the target medium is the single-differential cross section with respect to the energy transfer E in a collision. In this section, we discuss the calculation of $\mathrm{d}\sigma/\mathrm{d}E$ and its moments for "fast", point-like particles. To be precise, we consider particles with a velocity that is large compared to the velocities of the atomic electrons, corresponding to the domain of validity of the first-order Born approximation.

In the limit where the energy transfer E is large compared to the atomic binding energies, $\mathrm{d}\sigma/\mathrm{d}E$ approaches the cross section for scattering off a free electron. For a spin-zero particle with charge ze and speed βc, the asymptotic cross section (per electron) towards large energy transfers is given by [8]

$$\frac{\mathrm{d}\sigma}{\mathrm{d}E} = \underbrace{\frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2}\frac{1}{E^2}}_{\text{Rutherford cross section}}\left(1-\beta^2\frac{E}{E_{\max}}\right) = \frac{\mathrm{d}\sigma_R}{\mathrm{d}E}\left(1-\beta^2\frac{E}{E_{\max}}\right). \tag{2.11}$$

Similar expressions have been derived for particles with spin 1 and spin 1/2 [8]. The maximum energy transfer is given by the kinematics of a head-on collision between a particle with mass M and an electron (mass m) at rest,

$$E_{\max} = 2mc^2\beta^2\gamma^2\left(1+2\gamma\frac{m}{M}+\left(\frac{m}{M}\right)^2\right)^{-1}, \tag{2.12}$$

which for $M \gg m$ becomes $E_{\max} \sim 2mc^2\beta^2\gamma^2$.

These so-called "close" or "knock-on" collisions, in which the projectile interacts with a single atomic electron, contribute a significant fraction (roughly half) to the average energy loss of a charged particle in matter but are rare compared to "distant" collisions in which the particle interacts with the atom as a whole or with a group of

atoms. For an accurate calculation of $\mathrm{d}\sigma/\mathrm{d}E$, the electronic structure of the target medium therefore needs to be taken into account.

In the non-relativistic first-order Born approximation, the transition of an atom from its ground state to an excited state $|n\rangle$ involving a momentum transfer $\mathbf{q}$ is characterised by the matrix element (inelastic form factor)

$$F_{n0}(\mathbf{q}) = \langle n| \sum_{j=1}^{Z} \exp\left(\frac{\mathrm{i}}{\hbar}\mathbf{q}\cdot\mathbf{r}_j\right)|0\rangle,$$

which is independent of the projectile. The differential cross section with respect to the recoil parameter $Q = q^2/(2m)$, derived by Bethe in 1930 [16], is given by [1–3, 16]

$$\frac{\mathrm{d}\sigma_n}{\mathrm{d}Q} = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2}\frac{1}{Q^2}|F_{n0}(\mathbf{q})|^2 = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2}\frac{f_n(q)}{QE_n},$$

where $f_n(q)$ denotes the generalised oscillator strength (GOS). In the limit $q \to 0$ it becomes the dipole oscillator strength f_n discussed in Sect. 2.2.1. The double-differential cross section for transitions to the continuum (i.e. ionisation) is given by

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}E\mathrm{d}Q} = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2}\frac{1}{QE}\frac{\mathrm{d}f(E,q)}{\mathrm{d}E}, \tag{2.13}$$

where $\mathrm{d}f(E,q)/\mathrm{d}E$ is the generalised oscillator strength density. The GOS is constrained by the Bethe sum rule [2, 16] (a generalisation of the TRK sum rule),

$$\sum_n f_n(q) + \int \mathrm{d}E \frac{\mathrm{d}f(E,q)}{\mathrm{d}E} = Z, \quad \forall q. \tag{2.14}$$

Closed-form expressions for the generalised oscillator strength (density) exist only for very simple systems such as the hydrogen atom (Fig. 2.4). Numerical calculations are available for a number of atoms and molecules (see e.g. Ref. [37]). A prominent feature of the generalised oscillator strength density is the so-called "Bethe ridge": at high momentum transfers $\mathrm{d}f(E,q)/\mathrm{d}E$ is concentrated along the free-electron dispersion relation $Q = E$.

In order to calculate $\mathrm{d}\sigma/\mathrm{d}E$, we need to integrate the double-differential cross-section over Q,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}E} = \int_{Q_{\min}}^{Q_{\max}} \mathrm{d}Q \frac{\mathrm{d}^2\sigma}{\mathrm{d}E\mathrm{d}Q}, \qquad Q_{\min} \sim \frac{E^2}{2m\beta^2c^2}. \tag{2.15}$$

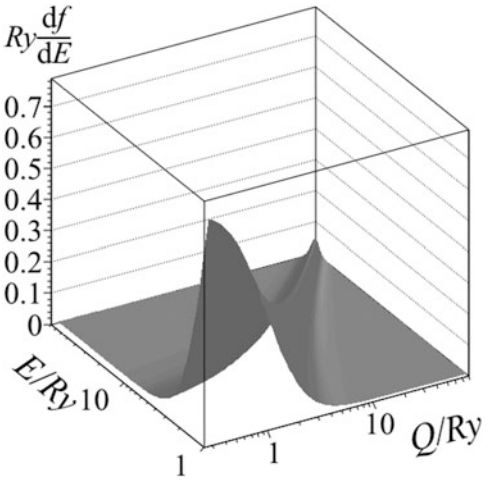


Fig. 2.4 Generalised oscillator strength density $\mathrm{d}f\,(E,q)\,/\mathrm{d}E$ of atomic hydrogen [2, 3, 16], for transitions to the continuum

For this purpose, it is often sufficient to use simplified models of the generalised oscillator strength density, based on the guidelines provided by model systems like the hydrogen atom, and using (measured) optical data in the low-Q regime.

Equation (2.13) describes the interaction of a charged particle with an isolated atom, which is a suitable approximation for a dilute gas. In order to extend it to dense media and to incorporate relativistic effects, it is convenient to use a semi-classical formalism [19, 38]. In this approach, which can be shown to be equivalent to the first-order quantum mechanical result, the response of the medium to the incident particle is described in terms of the complex dielectric function.

2.3.1 Dielectric Theory

Revisiting the energy loss of charged particles in matter from the viewpoint of classical electrodynamics, we calculate the electric field of a point charge ze moving with a constant velocity $\boldsymbol{\beta}c$ through an infinite, homogeneous and isotropic medium, that is we solve Maxwell's equations

$$\nabla\cdot\mathbf{B}=0\,,\qquad \nabla\times\mathbf{E}=-\frac{1}{c}\frac{\partial\mathbf{B}}{\partial t},$$
$$\nabla\times\mathbf{B}=\frac{1}{c}\frac{\partial\mathbf{D}}{\partial t}+\frac{4\pi}{c}\mathbf{j}\,,\qquad \nabla\cdot\mathbf{D}=4\pi\rho,$$

for source terms

$$\rho=ze\delta^3\left(\mathbf{r}-\boldsymbol{\beta}ct\right),\qquad \mathbf{j}=\boldsymbol{\beta}c\rho.$$

The perturbation due to the moving charge is assumed to be weak enough such that there is a linear relationship between the Fourier components of the electric field $\mathbf{E}$ and the displacement field $\mathbf{D}$,

$$\mathbf{D}(\mathbf{k},\omega) = \varepsilon(\mathbf{k},\omega)\,\mathbf{E}(\mathbf{k},\omega),$$

where $\varepsilon(\mathbf{k},\omega) = \varepsilon_1(\mathbf{k},\omega) + \mathrm{i}\varepsilon_2(\mathbf{k},\omega)$ is the (generalized) complex dielectric function.

The particle experiences a force $ze\mathbf{E}(\mathbf{r} = \boldsymbol{\beta}ct, t)$ that slows it down, and the stopping power is given by the component of this force parallel to the particle's direction of motion,

$$\frac{\mathrm{d}E}{\mathrm{d}x} = ze\mathbf{E}\cdot\frac{\boldsymbol{\beta}}{\beta}.$$

Adopting the Coulomb gauge $\mathbf{k}\cdot\mathbf{A} = 0$, one obtains after integrating over the angles (assuming that the dielectric function ε is isotropic),

$$\frac{\mathrm{d}E}{\mathrm{d}x} = -\frac{2z^2e^2}{\beta^2\pi}\int \mathrm{d}\omega \int \mathrm{d}k \times\left[\frac{\omega}{kc^2}\mathrm{Im}\left(\frac{-1}{\varepsilon(k,\omega)}\right) + \omega k\left(\beta^2 - \frac{\omega^2}{k^2c^2}\right)\mathrm{Im}\left(\frac{1}{-k^2c^2 + \varepsilon(k,\omega)\,\omega^2}\right)\right]. \tag{2.16}$$

The first term in the integrand represents the non-relativistic contribution to the energy loss which we would have obtained by considering only the scalar potential ϕ. It is often referred to as the longitudinal term. The second term, known as the transverse term, originates from the vector potential $\mathbf{A}$ and incorporates relativistic effects.

On a microscopic level, the energy transfer from the particle to the target medium proceeds through discrete collisions with energy transfer $E = \hbar\omega$ and momentum transfer $q = \hbar k$. Comparing Eq. (2.2) with the macroscopic result (2.16), one obtains

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}E\mathrm{d}q} = \frac{2z^2\alpha}{\beta^2\pi\hbar cN} \times\left[\frac{1}{q}\mathrm{Im}\left(\frac{-1}{\varepsilon(q,E)}\right) + \frac{1}{q}\left(\beta^2 - \frac{E^2}{q^2c^2}\right)\mathrm{Im}\left(\frac{1}{-1 + \varepsilon(q,E)\,E^2/\left(q^2c^2\right)}\right)\right]. \tag{2.17}$$

The loss function $\mathrm{Im}\left(-1/\varepsilon\left(q,E\right)\right)$ and the generalized oscillator strength density are related by

$$\frac{\mathrm{d}f\left(E,q\right)}{\mathrm{d}E}=E\frac{2Z}{\pi\left(\hbar\Omega_p\right)^2}\mathrm{Im}\left(\frac{-1}{\varepsilon\left(q,E\right)}\right). \tag{2.18}$$

Using this identity, we see that the longitudinal term (first term) in Eq. (2.17) is equivalent to the non-relativistic quantum mechanical result (2.13). As is the case with the generalized oscillator strength density, closed-form expressions for the dielectric loss function $\mathrm{Im}\left(-1/\varepsilon\left(q,E\right)\right)$ can only be derived for simple systems like the ideal Fermi gas [39, 40]. In the following (Sects. 2.3.2 and 2.3.3), we discuss two specific models of $\mathrm{Im}\left(-1/\varepsilon\left(q,E\right)\right)$ (or, equivalently, $\mathrm{d}f\left(E,q\right)/\mathrm{d}E$).

2.3.2 Bethe-Fano Method

The relativistic version of Eq. (2.13) or, in other words, the equivalent of Eq. (2.17) in oscillator strength parlance, is [1, 41]

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}E\mathrm{d}Q}=\frac{2\pi z^2\left(\alpha\hbar c\right)^2}{mc^2\beta^2}Z\left[\frac{\left|F\left(E,\mathbf{q}\right)\right|^2}{Q^2\left(1+\frac{Q}{2mc^2}\right)^2}+\frac{\left|\boldsymbol{\beta}_t\cdot\mathbf{G}\left(E,\mathbf{q}\right)\right|^2}{\left[Q\left(1+\frac{Q}{2mc^2}\right)-\frac{E^2}{2mc^2}\right]^2}\right]\left(1+\frac{Q}{mc^2}\right) \tag{2.19}$$

where $Q\left(1+Q/2mc^2\right)=q^2/2m$, $\boldsymbol{\beta}_t$ is the component of the velocity perpendicular to the momentum transfer $\mathbf{q}$, and $F\left(E,\mathbf{q}\right)$ and $\mathbf{G}\left(E,\mathbf{q}\right)$ represent the matrix elements for longitudinal and transverse excitations.

Depending on the type of target and the range of momentum transfers involved, we can use Eqs. (2.13), (2.19) or (2.17) as a starting point for evaluating the single-differential cross section. Following the approach described by Fano [1], we split $\mathrm{d}\sigma/\mathrm{d}E$ in four parts. For small momentum transfers ($Q<Q_1\sim 1\,\mathrm{Ry}$), we can use the non-relativistic expression (2.13) for the longitudinal term and approximate the generalised oscillator strength density by its dipole limit,

$$\frac{\mathrm{d}\sigma^{(1)}}{\mathrm{d}E}=\frac{2\pi z^2\left(\alpha\hbar c\right)^2}{mc^2\beta^2}\frac{1}{E}\frac{\mathrm{d}f\left(E\right)}{\mathrm{d}E}\int_{Q_{\min}}^{Q_1}\frac{\mathrm{d}Q}{Q}=\frac{2\pi z^2\left(\alpha\hbar c\right)^2}{mc^2\beta^2}\frac{1}{E}\frac{\mathrm{d}f\left(E\right)}{\mathrm{d}E}\ln\frac{Q_1 2mc^2\beta^2}{E^2}. \tag{2.20}$$

In terms of the dielectric loss function, one obtains

$$\frac{\mathrm{d}\sigma^{(1)}}{\mathrm{d}E}=\frac{z^2\alpha}{\beta^2\pi\hbar cN}\mathrm{Im}\left(\frac{-1}{\varepsilon\left(E\right)}\right)\ln\frac{Q_1 2mc^2\beta^2}{E^2}.$$

For high momentum transfers ($Q > Q_2 \sim 30\,\mathrm{keV}$), i.e. for close collisions where the binding energy of the atomic electrons can be neglected, the longitudinal and transverse matrix elements are strongly peaked at the Bethe ridge $Q = E$. Using [1]

$$|F(E,\mathbf{q})|^2 \sim \frac{1 + Q/\left(2mc^2\right)}{1 + Q/\left(mc^2\right)}\delta(E-Q),$$

$$|\boldsymbol{\beta}_t \cdot \mathbf{G}(E,\mathbf{q})|^2 \sim \beta_t^2 \frac{1 + Q/\left(2mc^2\right)}{1 + Q/\left(mc^2\right)}\delta(E-Q)$$

and

$$\beta_t^2 = \frac{1}{1 + Q/\left(2mc^2\right)} - \left(1-\beta^2\right)$$

one obtains (for longitudinal and transverse excitations combined),

$$\frac{\mathrm{d}\sigma^{(h)}}{\mathrm{d}E} = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2}\frac{Z}{E}\left(1 - \frac{E\left(1-\beta^2\right)}{2mc^2}\right). \tag{2.21}$$

In the intermediate range, $Q_1 < Q < Q_2$, numerical calculations of the generalised oscillator strength density are used. An example of $\mathrm{d}f(E,q)/\mathrm{d}E$ is shown in Fig. 2.5. Since the limits Q_1, Q_2 do not depend on the particle velocity, the integrals

$$\frac{\mathrm{d}\sigma^{(2)}}{\mathrm{d}E} = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2}\frac{1}{E}\int_{Q_1}^{Q_2}\frac{\mathrm{d}Q}{Q}\frac{\mathrm{d}f(E,q)}{\mathrm{d}E}$$

need to be evaluated only once for each value of E. The transverse contribution can be neglected[1] [41].

The last contribution, described in detail in Ref. [1], is due to low-Q transverse excitations in condensed matter. Setting $\mathrm{Im}\left(-1/\varepsilon(E,q)\right) = \mathrm{Im}\left(-1/\varepsilon(E)\right)$ in the second term in Eq. (2.17) and integrating over q gives

$$\frac{\mathrm{d}\sigma^{(3)}}{\mathrm{d}E} = \frac{z^2\alpha}{\beta^2\pi N\hbar c} \times\left[\mathrm{Im}\left(\frac{-1}{\varepsilon(E)}\right)\ln\frac{1}{\left|1-\beta^2\varepsilon(E)\right|} + \left(\beta^2 - \frac{\varepsilon_1(E)}{|\varepsilon(E)|^2}\right)\left(\frac{\pi}{2} - \arctan\frac{1-\beta^2\varepsilon_1(E)}{\beta^2\varepsilon_2(E)}\right)\right]. \tag{2.22}$$

[1] For particle speeds $\beta < 0.1$, this approximation will cause errors, especially for M_0.

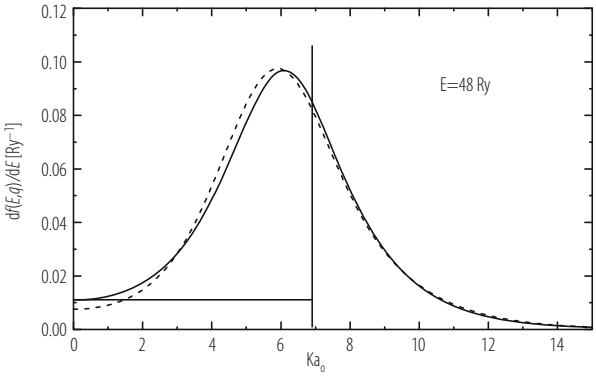


Fig. 2.5 Generalized oscillator strength density for Si for an energy transfer $E = 48\,\mathrm{Ry}$ to the 2p-shell electrons [41–44], as function of ka_0 (where $k^2a_0^2 = Q/\mathrm{Ry}$). Solid line: calculated with Herman-Skilman potential, dashed line: hydrogenic approximation [45, 46]. The horizontal and vertical line define the FVP approximation (Sect. 2.3.3)

We will discuss this term in more detail in Sect. 2.3.3. The total single-differential cross section,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}E} = \frac{\mathrm{d}\sigma^{(1)}}{\mathrm{d}E} + \frac{\mathrm{d}\sigma^{(2)}}{\mathrm{d}E} + \frac{\mathrm{d}\sigma^{(3)}}{\mathrm{d}E} + \frac{\mathrm{d}\sigma^{(h)}}{\mathrm{d}E},$$

is shown in Fig. 2.6 for particles with $\beta\gamma = 4$ in silicon which, at present, is the only material for which calculations based on the Bethe-Fano method are available.

2.3.3 *Fermi Virtual-Photon (FVP) Method*

In the Bethe-Fano algorithm discussed in the previous section, the dielectric function $\varepsilon\,(q, E)$ was approximated at low momentum transfer by its optical limit $\varepsilon\,(E)$. In the Fermi virtual-photon (FVP) or Photoabsorption Ionisation (PAI) model [6, 47, 48], this approximation is extended to the entire domain $q^2 < 2mE$. Guided by the shape of the hydrogenic GOS, the remaining contribution to $\mathrm{Im}\,(-1/\varepsilon\,(q, E))$ required to satisfy the Bethe sum rule

$$\int_0^\infty E\,\mathrm{Im}\left(\frac{-1}{\varepsilon\,(q, E)}\right)\mathrm{d}E = \frac{\pi}{2}\left(\hbar\Omega_p\right)^2 \quad \forall q, \tag{2.23}$$

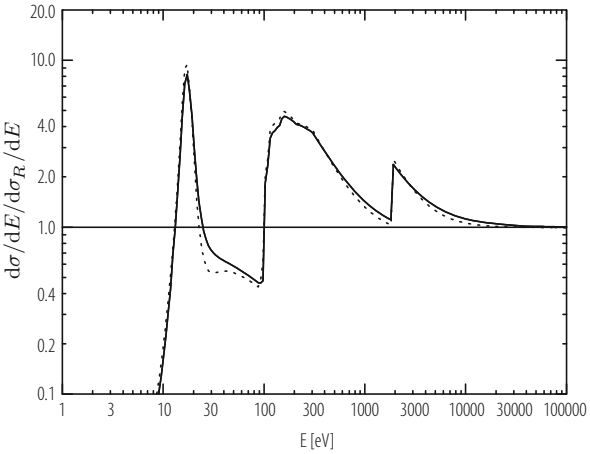


Fig. 2.6 Differential cross section $\mathrm{d}\sigma/\mathrm{d}E$, divided by the Rutherford cross section $\mathrm{d}\sigma_R/\mathrm{d}E$, for particles with $\beta\gamma = 4$ in silicon, calculated with two methods. The abscissa is the energy loss E in a single collision. The Rutherford cross section is represented by the horizontal line at 1.0. The solid line was obtained [41] with the Bethe-Fano method (Sect. 2.3.2). The cross section calculated with the FVP method (Sect. 2.3.3) is shown by the dotted line. The functions all extend to $E_{\max} \sim 16\,\mathrm{MeV}$. The moments are $M_0 = 4$ collisions/μm and $M_1 = 386\,\mathrm{eV/\mu m}$ (Table 2.2)

is attributed to the scattering off free electrons (close collisions). This term is thus of the form $C\delta\left(E - q^2/(2m)\right)$, with the factor C being determined by the normalisation (2.23),

$$C = \frac{1}{E}\int_0^E E'\mathrm{Im}\left(\frac{-1}{\varepsilon\left(E'\right)}\right)\mathrm{d}E'.$$

Combining the two terms, the longitudinal loss function becomes

$$\mathrm{Im}\left(\frac{-1}{\varepsilon\left(q,\,E\right)}\right) = \mathrm{Im}\left(\frac{-1}{\varepsilon\left(E\right)}\right)\Theta\left(E - \frac{q^2}{2m}\right) + \frac{\delta\left(E - \frac{q^2}{2m}\right)}{E}\int_0^E E'\mathrm{Im}\left(\frac{-1}{\varepsilon\left(E'\right)}\right)\mathrm{d}E'.$$

In the transverse term, the largest contribution to the integral comes from the region $E \sim qc/\sqrt{\varepsilon}$, i.e. from the vicinity of the (real) photon dispersion relation, and one consequently approximates $\varepsilon\left(q,\,E\right)$ by $\varepsilon\left(E\right)$ throughout.

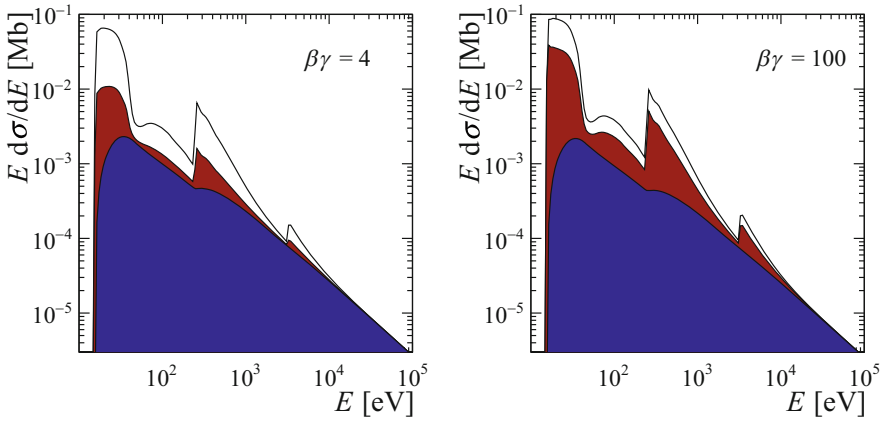


Fig. 2.7 Differential cross section $\mathrm{d}\sigma/\mathrm{d}E$ (scaled by the energy loss E) calculated using the FVP algorithm, for particles with $\beta\gamma = 4$ (left) and $\beta\gamma = 100$ (right) in argon (at atmospheric pressure, $T = 20\,^{\circ}\mathrm{C}$). The upper, unshaded area corresponds to the first term in Eq. (2.24), i.e. to the contribution from distant longitudinal collisions. The lower area corresponds to the contribution from close longitudinal collisions, given by the second term in Eq. (2.24). The intermediate area corresponds to the contribution from transverse collisions

The integration over q can then be carried out analytically and one obtains for the single-differential cross section $\mathrm{d}\sigma/\mathrm{d}E$

$$\frac{\mathrm{d}\sigma}{\mathrm{d}E} = \frac{z^2\alpha}{\beta^2\pi N\hbar c}\left[\mathrm{Im}\left(\frac{-1}{\varepsilon\,(E)}\right)\ln\frac{2mc^2\beta^2}{E} + \frac{1}{E^2}\int_0^E E'\mathrm{Im}\left(\frac{-1}{\varepsilon\,(E')}\right)\mathrm{d}E'\right] + \frac{z^2\alpha}{\beta^2\pi N\hbar c}$$
$$\times\left[\mathrm{Im}\left(\frac{-1}{\varepsilon\,(E)}\right)\ln\frac{1}{\left|1-\beta^2\varepsilon\,(E)\right|} + \left(\beta^2 - \frac{\varepsilon_1\,(E)}{|\varepsilon\,(E)|^2}\right)\left(\frac{\pi}{2} - \arctan\frac{1-\beta^2\varepsilon_1\,(E)}{\beta^2\varepsilon_2\,(E)}\right)\right] \tag{2.24}$$

The relative importance of the different terms in Eq. (2.24) is illustrated in Fig. 2.7. The first two terms describe the contributions from longitudinal distant and close collisions. The contribution from transverse collisions (third and fourth term) is identical to $\mathrm{d}\sigma^{(3)}/\mathrm{d}E$ in the Bethe-Fano algorithm. As can be seen from Fig. 2.7, its importance grows with increasing $\beta\gamma$. The third term incorporates the relativistic density effect, i.e. the screening of the electric field due to the polarisation of the medium induced by the passage of the charged particle. In the transparency region $\varepsilon_2\,(E) = 0$, the fourth term can be identified with the cross section for the emission of Cherenkov photons. It vanishes for $\beta < 1/\sqrt{\varepsilon}$; above this threshold it becomes

$$\frac{\mathrm{d}\sigma^{(\mathrm{C})}}{\mathrm{d}E} = \frac{\alpha}{N\hbar c}\left(1 - \frac{1}{\beta^2\varepsilon}\right) \sim \frac{\alpha}{N\hbar c}\sin^2\theta_{\mathrm{C}},$$

where

$$\cos\theta_{\mathrm{C}} = \frac{1}{\beta\sqrt{\varepsilon}}.$$

Cherenkov detectors are discussed in detail in Chap. 7 of this book.

In the formulation of the PAI model by Allison and Cobb [6], the imaginary part ε_2 of the dielectric function is approximated by the photoabsorption cross section σ_γ,

$$\varepsilon_2(E) \sim \frac{N\hbar c}{E}\sigma_\gamma(E) \tag{2.25}$$

and the real part ε_1 is calculated from the Kramers-Kronig relation

$$\varepsilon_1(E) - 1 = \frac{2}{\pi}\mathsf{P}\int_0^\infty \frac{E'\varepsilon_2(E')}{E'^2 - E^2}\mathrm{d}E'.$$

In addition, the approximation $|\varepsilon(E)|^2 \sim 1$ is used. These are valid approximations if the refractive index[2] is close to one ($\mathsf{n} \sim 1$) and the attenuation coefficient k is small. For gases, this requirement is usually fulfilled for energies above the ionisation threshold.

Requiring only optical data as input, the FVP/PAI model is straightforward to implement in computer simulations. In the HEED program [49], the differential cross section $\mathrm{d}\sigma/\mathrm{d}E$ is split into contributions from each atomic shell, which enables one to simulate not only the energy transfer from the projectile to the medium but also the subsequent atomic relaxation processes (Sect. 2.6). The GEANT4 implementation of the PAI model is described in Ref. [50]. For reasons of computational efficiency, the photoabsorption cross section $\sigma_\gamma(E)$ is parameterised as a fourth-order polynomial in $1/E$. FVP calculations for Ne and Ar/CH_4 (90:10) are discussed in Ref. [51].

2.3.4 *Integral Quantities*

For validating and comparing calculations of the differential cross section, it is instructive to consider the moments M_i of $N\mathrm{d}\sigma/\mathrm{d}E$, in particular the inverse mean free path M_0 and the stopping power M_1.

[2]The complex refractive index and the dielectric function are related by $\mathsf{n} + \mathrm{i}\mathsf{k} = \sqrt{\varepsilon}$.

2.3.4.1 Inverse Mean Free Path

In the relativistic first-order Born approximation, the inverse mean free path for ionising collisions has the form [1, 2]

$$M_0 = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2} N \left[M^2 \left(\ln\left(\beta^2\gamma^2\right) - \beta^2 \right) + C \right], \tag{2.26}$$

where

$$M^2 = \int \frac{1}{E} \frac{\mathrm{d}f\,(E)}{\mathrm{d}E} \mathrm{d}E, \qquad C = M^2 \left(\ln \tilde{c} + \ln \frac{4}{\alpha^2} \right),$$

and $\tilde{c}$ is a material-dependent parameter that can be calculated from the generalised oscillator strength density. Calculations can be found, for example, in Refs. [53, 54]. As in the Bethe stopping formula (2.28) discussed below, a correction term can be added to Eq. (2.26) to account for the density effect [55].

The inverse mean free path for dipole-allowed discrete excitations is given by [2]

$$M_0^{(n)} = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2} N \frac{f_n}{E_n} \left[\ln\left(\beta^2\gamma^2\right) - \beta^2 + \ln \tilde{c}_n + \ln \frac{4}{\alpha^2} \right].$$

We can thus obtain a rough estimate of the relative frequencies of excitations and ionising collisions from optical data. In argon, for instance, the ratio of $\sum f_n/E_n$ and M^2 is ∼20% [25].

For gases, M_0 can be determined experimentally by measuring the inefficiency of a gas-filled counter operated at high gain ("zero-counting method"). Results (in the form of fit parameters M^2, C) from an extensive series of measurements, using electrons with kinetic energies between 0.1 and 2.7 MeV, are reported in Ref. [52]. Other sets of experimental data obtained using the same technique can be found in Refs. [56, 57]. Table 2.1 shows a comparison between measured and calculated values (using the FVP algorithm) of M_0 for particles with $\beta\gamma = 3.5$ at a temperature of 20 °C and atmospheric pressure. The inverse mean free path is

Table 2.1 Measurements [52] and calculations (using the FVP algorithm as implemented in HEED [49]) of M_0 for $\beta\gamma = 3.5$ at $T = 20\,^{\circ}\mathrm{C}$ and atmospheric pressure

	M_0 [cm^{-1}]	
Gas	Measurement	FVP
Ne	10.8	10.5
Ar	23.0	25.4
Kr	31.5	31.0
Xe	43.2	42.1
CO_2	34.0	34.0
CF_4	50.9	51.8
CH_4	24.6	29.4
iC_4H_{10}	83.4	90.9

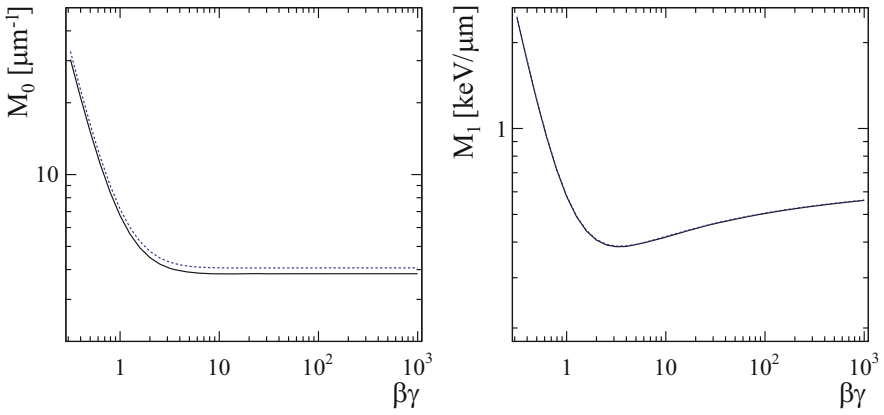


Fig. 2.8 Inverse ionisation mean free path (left) and stopping power (right) of heavy charged particles in silicon as a function of $\beta\gamma$, calculated using the Bethe-Fano algorithm (solid line) and the FVP model (dashed line). The two stopping power curves are virtually identical

sensitive to the detailed shape of the differential cross section $\mathrm{d}\sigma/\mathrm{d}E$ at low energies and, consequently, to the optical data used.

Figure 2.8(left) shows M_0 in solid silicon as a function of $\beta\gamma$, calculated using the Bethe-Fano and FVP algorithms. The difference between the results is $\sim 6-8\%$, as can also be seen from Table 2.2. Owing to the more detailed (and more realistic) modelling of the generalised oscillator strength density at intermediate Q, the Bethe-Fano algorithm can be expected to be more accurate than the FVP method.

2.3.4.2 Stopping Power

Let us first consider the average energy loss of a non-relativistic charged particle in a dilute gas, with the double-differential cross section given by Eq. (2.13),

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2} N \int_{E_{\min}}^{E_{\max}} \mathrm{d}E \int_{Q_{\min}}^{Q_{\max}} \frac{\mathrm{d}Q}{Q} \frac{\mathrm{d}f(E,q)}{\mathrm{d}E}.$$

As an approximation, we assume that the integrations over Q and E can be interchanged and the integration limits $Q_{\min}, Q_{\max}$ (which depend on E) be replaced by average values $\overline{Q}_{\min} = I^2/\left(2m\beta^2c^2\right)$, $\overline{Q}_{\max} = E_{\max}$ [58]. Using the Bethe sum rule (2.23), we then obtain

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{2\pi z^2 (\alpha\hbar c)^2}{mc^2\beta^2} NZ \ln \frac{2mc^2\beta^2 E_{\max}}{I^2},$$

where the target medium is characterised by a single parameter: the "mean ionisation energy" I, defined by

$$\ln I = \frac{1}{Z}\int \mathrm{d}E \ln E \frac{\mathrm{d}f\,(E)}{\mathrm{d}E}$$

in terms of the dipole oscillator strength density, or

$$\ln I = \frac{2}{\pi\left(\hbar\Omega_p\right)^2}\int \mathrm{d}E\; E\,\mathrm{Im}\left(\frac{-1}{\varepsilon\,(E)}\right)\ln E. \tag{2.27}$$

in terms of the dielectric loss function.

In the relativistic case, one finds the well-known Bethe stopping formula

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{2\pi z^2\,(\alpha\hbar c)^2}{mc^2\beta^2} NZ\left[\ln\frac{2mc^2\beta^2\gamma^2 E_{\max}}{I^2} - 2\beta^2 - \delta\right], \tag{2.28}$$

where δ is a correction term accounting for the density effect [59].

Sets of stopping power tables for protons and alpha particles are available in ICRU report 49 [60] and in the PSTAR and ASTAR online databases [61]. Tables for muons are given in Ref. [62]. These tabulations include stopping power contributions beyond the first-order Born approximation, such as shell corrections [42, 45, 46] and the Barkas-Andersen effect [63–65].

The stopping power in silicon obtained from the Bethe-Fano algorithm (Sect. 2.3.2) has been found to agree with measurements within ±0.5% [41]. As can be seen from Table 2.2 and Fig. 2.8, FVP and Bethe-Fano calculations for M_1 in silicon are in close agreement, with differences <1%.

In addition to M_0, M_1, Table 2.2 also includes the most probable value of the energy loss spectrum in an 8 µm thick layer of silicon. For thin absorbers, as will be discussed in Sect. 2.5, the stopping power $\mathrm{d}E/\mathrm{d}x$ is not a particularly meaningful quantity for characterising energy loss spectra. Because of the asymmetric shape of the differential cross section $\mathrm{d}\sigma/\mathrm{d}E$, the most probable value Δ_p of the energy loss distribution is typically significantly smaller than the average energy loss $\langle\Delta\rangle = M_1 x$.

2.4 Electron Collisions and Bremsstrahlung

The formalism for computing the differential cross section $\mathrm{d}\sigma/\mathrm{d}E$ for collisions of heavy charged particles with the electrons of the target medium, discussed in Sect. 2.3, is also applicable to electron and positron projectiles, except that the asymptotic close-collision cross section (2.11) is replaced by the Møller and Bhabha cross sections respectively [8, 66]. When evaluating the inverse mean free path M_0 or the stopping power M_1, we further have to take into account that the energy loss

Table 2.2 Integral properties of collision cross sections for Si calculated with Bethe-Fano (B-F) and FVP algorithms

	M_0 [μm^{-1}]		M_1 [eV/μm]		Δ_p/x [eV/μm]	
$\beta\gamma$	B-F	FVP	B-F	FVP	B-F	FVP
0.316	30.325	32.780	2443.72	2465.31	1677.93	1722.92
0.398	21.150	22.781	1731.66	1745.57	1104.90	1135.68
0.501	15.066	16.177	1250.93	1260.18	744.60	765.95
0.631	11.056	11.840	928.70	935.08	520.73	536.51
0.794	8.433	9.010	716.37	720.98	381.51	394.03
1.000	6.729	7.175	578.29	581.79	294.54	304.89
1.259	5.632	5.996	490.84	493.65	240.34	249.25
1.585	4.932	5.245	437.34	439.72	207.15	215.02
1.995	4.492	4.771	406.59	408.70	187.39	194.60
2.512	4.218	4.476	390.95	392.89	176.30	183.06
3.162	4.051	4.296	385.29	387.12	170.70	177.16
3.981	3.952	4.189	386.12	387.89	168.59	174.81
5.012	3.895	4.127	391.08	392.80	168.54	174.63
6.310	3.865	4.094	398.54	400.24	169.62	175.60
7.943	3.849	4.076	407.39	409.07	171.19	177.10
10.000	3.842	4.068	416.91	418.58	172.80	178.66
12.589	3.839	4.064	426.63	428.29	174.26	180.06
15.849	3.839	4.063	436.30	437.96	175.45	181.24
19.953	3.839	4.063	445.79	447.44	176.36	182.14
25.119	3.840	4.063	455.03	456.68	177.04	182.79
31.623	3.840	4.064	463.97	465.63	177.53	183.28
39.811	3.841	4.064	472.61	474.27	177.86	183.61
50.119	3.842	4.065	480.93	482.58	178.09	183.83
63.096	3.842	4.065	488.90	490.55	178.22	183.95
79.433	3.842	4.065	496.52	498.17	178.32	184.06
100.000	3.842	4.066	503.77	505.42	178.38	184.10
125.893	3.843	4.066	510.66	512.31	178.43	184.15
158.489	3.843	4.066	517.20	518.84	178.44	184.17
199.526	3.843	4.066	523.40	525.05	178.47	184.18
251.189	3.843	4.066	529.29	530.94	178.48	184.18
316.228	3.843	4.066	534.91	536.56	178.48	184.21
398.107	3.843	4.066	540.28	541.92	178.48	184.22
501.187	3.843	4.066	545.43	547.08	178.48	184.22
630.958	3.843	4.066	550.40	552.05	178.48	184.22
794.329	3.843	4.066	555.21	556.86	178.48	184.22
1000.000	3.843	4.066	559.89	561.54	178.48	184.22

The third column shows the most probable value Δ_p of the energy loss spectrum divided by the track length x, for $x = 8\,\mu m$. The minimum values for M_0 are at $\beta\gamma \sim 18$, for M_1 at $\beta\gamma \sim 3.2$, for Δ_p at $\beta\gamma \sim 5$. The relativistic rise for M_0 is 0.1%, for M_1 it is 45%, for Δ_p it is 6%

of an electron in an ionising collision is limited to half of its kinetic energy,

$$E_{\max} = \frac{1}{2} mc^2 (\gamma - 1) , \tag{2.29}$$

as primary and secondary electron are indistinguishable. Stopping power tables for electrons are available in ICRU report 37 [67] and in the ESTAR database [61].

The other main mechanism by which fast electrons and positrons lose energy when traversing matter is the emission of radiation (bremsstrahlung) due to deflections in the electric field of the nucleus and the atomic electrons.

2.4.1 Bremsstrahlung

Let us first consider electron-nucleus bremsstrahlung, the first quantum-mechanical description of which was developed by Bethe and Heitler [68]. The differential cross section (per atom) for the production of a bremsstrahlung photon of energy E by an incident electron of kinetic energy T is given by [8, 68]

$$\frac{\mathrm{d}\sigma_{\mathrm{rad}}}{\mathrm{d}E} = 4\alpha^3 \left(\frac{\hbar c}{mc^2} \right)^2 Z^2 \frac{F(u, T)}{E}, \tag{2.30}$$

where $u = E/\left(\gamma mc^2\right)$ denotes the ratio of the photon energy to the projectile energy. Expressions for the function $F(u, T)$ are reviewed in Ref. [69] and can be fairly complex. Amongst other parameters, $F(u, T)$ depends on the extent to which the charge of the nucleus is screened by the atomic electrons. In the first-order Born approximation and in the limit of complete screening, applicable at high projectile energies, one obtains [8, 68, 69]

$$F(u) = \left(1 + (1-u)^2 - \frac{2}{3}(1-u) \right) \ln \frac{183}{Z^{1/3}} + \frac{1}{9}(1-u) . \tag{2.31}$$

The theoretical description of electron-electron bremsstrahlung is similar to the electron-nucleus case, except that the differential cross section is proportional to Z instead of Z^2. To a good approximation, we can include electron-electron bremsstrahlung in Eq. (2.30) by replacing the factor Z^2 by $Z(Z+1)$.

The inverse mean free path for the emission of a bremsstrahlung photon with energy $E > E_{\mathrm{cut}}$ is given by

$$\lambda^{-1} = M_0 = N \int_{E_{\mathrm{cut}}}^{T} \frac{\mathrm{d}\sigma_{\mathrm{rad}}}{\mathrm{d}E} \mathrm{d}E .$$

If we neglect the term $(1-u)/9$ in Eq. (2.31), we find for the radiative stopping power at $T \gg mc^2$

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = M_1 = N \int_0^T E \frac{\mathrm{d}\sigma_{\mathrm{rad}}}{\mathrm{d}E} \mathrm{d}E \sim \frac{T}{X_0}, \tag{2.32}$$

where the parameter X_0, defined by

$$\frac{1}{X_0} = 4\alpha^3 \left(\frac{\hbar c}{mc^2}\right)^2 NZ\,(Z+1) \ln \frac{183}{Z^{1/3}}, \tag{2.33}$$

is known as the radiation length. Values of X_0 for many commonly used materials can be found in Ref. [70] and on the PDG webpage [71]. Silicon, for instance, has a radiation length of $X_0 \sim 9.37\,\mathrm{cm}$ [71].

Being approximately proportional to the kinetic energy of the projectile, the radiative stopping power as a function of T increases faster than the average energy loss due to ionising collisions given by Eq. (2.28). At high energies—more precisely, above a so-called critical energy ($\sim$38 MeV in case of silicon [71])—bremsstrahlung therefore represents the dominant energy loss mechanism of electrons and positrons.

2.5 Energy Losses Along Tracks: Multiple Collisions and Spectra

Consider an initially monoenergetic beam of identical particles traversing a layer of material of thickness x. Due to the randomness both in the number of collisions and in the energy loss in each of the collisions, the total energy loss Δ in the absorber will vary from particle to particle. Depending on the use case, the kinetic energy of the particles, and the thickness x, different techniques for calculating the probability distribution $f(\Delta, x)$—known as "straggling function" [72]—can be used.

Our focus in this section is on scenarios where the average energy loss in the absorber is small compared to the kinetic energy T of the incident particle (as is usually the case in vertex and tracking detectors), such that the differential cross section $\mathrm{d}\sigma/\mathrm{d}E$ and its moments do not change significantly between the particle's entry and exit points in the absorber. The number of collisions n then follows a Poisson distribution

$$p(n, x) = \frac{\langle n \rangle^n}{n!} \mathrm{e}^{-\langle n \rangle}, \tag{2.34}$$

with mean $\langle n \rangle = x M_0$. The probability $f^{(1)}(E)\,\mathrm{d}E$ for a particle to lose an amount of energy between E and $E + \mathrm{d}E$ in a single collision is given by the normalised

differential cross section,

$$f^{(1)}(E) = \frac{1}{M_0} N \frac{\mathrm{d}\sigma}{\mathrm{d}E},$$

and the probability distribution for a total energy loss Δ in n collisions is obtained from n-fold convolution of $f^{(1)}$,

$$f^{(n)}(\Delta) = \underbrace{\left(f^{(1)} \otimes f^{(1)} \otimes \cdots \otimes f^{(1)}\right)}_{n \text{ times}}(\Delta) = \int \mathrm{d}E \; f^{(n-1)}(\Delta - E)\, f^{(1)}(E),$$

as illustrated in Figs. 2.9 and 2.10.

The probability distribution for a particle to suffer a total energy loss Δ over a fixed distance x is given by [72, 73]

$$f(\Delta, x) = \sum_{n=0}^{\infty} p(n, x)\, f^{(n)}(\Delta), \tag{2.35}$$

where $f^{(0)}(\Delta) = \delta(\Delta)$. Equation (2.35) can be evaluated in a stochastic manner (Sect. 2.5.1), by means of direct numerical integration (Sect. 2.5.2), or by using integral transforms (Sect. 2.5.3).

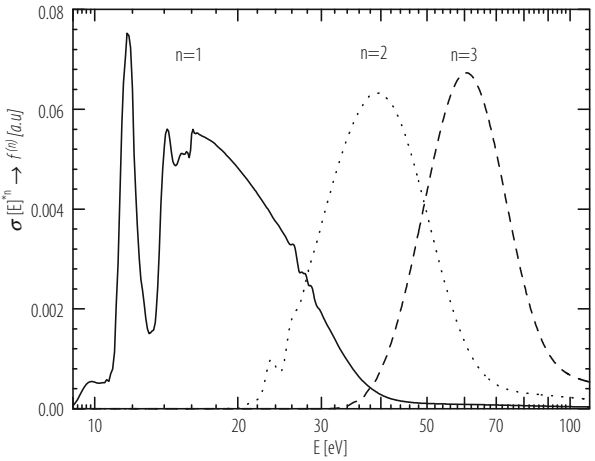


Fig. 2.9 Distributions $f^{(n)}$ of the energy loss in n collisions (n-fold convolution of the single-collision energy loss spectrum) for Ar/CH_4 (90:10)

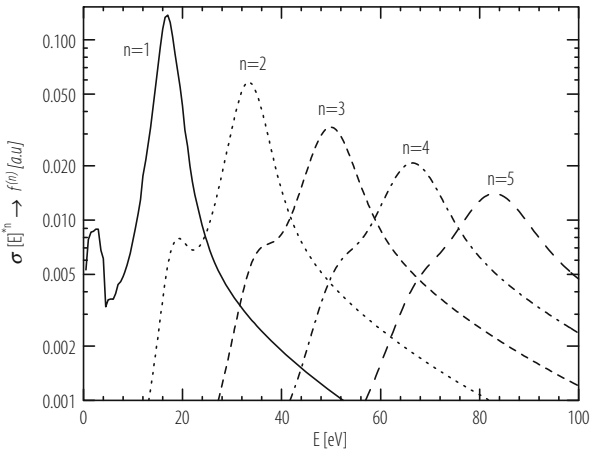


Fig. 2.10 Distributions $f^{(n)}$ of the energy loss in n collisions for solid silicon. The plasmon peak at ∼17 eV appears in each spectrum at $E \sim n \times 17$ eV, and its FWHM is proportional to $\sqrt{n}$. The structure at ∼2 eV appears at $2 + 17(n - 1)$ eV, but diminishes with increasing n. For $n = 6$ (not shown) the plasmon peak (at 102 eV) merges with the L-shell energy losses at 100 eV, also see Fig. 2.12

2.5.1 Monte Carlo Method

In a detailed Monte Carlo simulation, the trajectory of a single incident particle is followed from collision to collision. The required ingredients are the inverse mean free path $M_0^{(i)}$ and the cumulative distribution function,

$$\Phi^{(i)}(E) = \frac{1}{M_0^{(i)}} \int_0^E N \frac{\mathrm{d}\sigma^{(i)}}{\mathrm{d}E'} \mathrm{d}E', \tag{2.36}$$

for each interaction process i (electronic collisions, bremsstrahlung, etc.) to be taken into account in the simulation. The distance Δx between successive collisions follows an exponential distribution and is sampled according to

$$\Delta x = -\frac{\ln r}{\lambda^{-1}},$$

where $r \in (0, 1]$ is a uniformly distributed random number and

$$\lambda^{-1} = \sum_i M_0^{(i)}$$

is the total inverse mean free path. After updating the coordinates of the particle, the collision mechanism to take place is chosen based on the relative frequencies $M_0^{(i)}/\lambda^{-1}$. The energy loss in the collision is then sampled by drawing another uniform random variate $u \in [0, 1]$, and determining the corresponding energy loss E from the inverse of the cumulative distribution,

$$E = \Phi^{-1}(u) .$$

In general, the new direction after the collision will also have to be sampled from a suitable distribution. The above procedure is repeated until the particle has left the absorber. The spectrum $f(\Delta, x)$ is found by simulating a large number of particles and recording the energy loss Δ in a histogram. Advantages offered by the Monte Carlo approach include its straightforward implementation, the possibility of including interaction mechanisms other than inelastic scattering (bremsstrahlung, elastic scattering etc.), and the fact that it does not require approximations to the shape of $\mathrm{d}\sigma/\mathrm{d}E$ to be made.

For thick absorbers, detailed simulations can become unpractical due to the large number of collisions, and the need to update the inverse mean free path M_0 and the cumulative distribution $\Phi(E)$ following the change in velocity of the particle.

In "mixed" simulation schemes, a distinction is made between "hard" collisions which are simulated individually, and "soft" collisions (e.g. elastic collisions with a small angular deflection of the projectile, or emission of low-energy bremsstrahlung photons) the cumulative effect of which is taken into account after each hard scattering event. Details on the implementation of mixed Monte Carlo simulations can be found, for example, in the PENELOPE user guide [74].

2.5.2 Convolutions

For short track segments, one can calculate the distributions $f^{(n)}$ explicitly by numerical integration and construct $f(\Delta, x)$ directly using Eq. (2.35). A computationally more efficient approach is the absorber doubling method [41, 75], which proceeds as follows. Consider a step x that is small compared to the mean free path such that $\langle n \rangle \ll 1$ (in practice: $\langle n \rangle < 0.01$ [76]). Expanding Eq. (2.35) in powers of $\langle n \rangle$ and retaining only constant and linear terms gives

$$f(E, x) \sim (1 - \langle n \rangle) f^{(0)}(E) + \langle n \rangle f^{(1)}(E) .$$

The straggling function for a distance $2x$ is then calculated using

$$f(\Delta, 2x) = \int_0^{\Delta} f(\Delta - E, x) f(E, x) \, \mathrm{d}E .$$

This procedure is carried out k times until the desired thickness $2^k x$ is reached. Because of the tail of $f^{(1)}(E)$ towards large energy transfers, the numerical convolution is performed on a logarithmic grid. More details of the implementation can be found in Refs. [75, 76].

2.5.3 Laplace Transforms

In the Laplace domain, Eq. (2.35) becomes

$$F(s,x) = \mathcal{L}\{f(\Delta,x)\} = \mathrm{e}^{-\langle n\rangle}\sum_{n=0}^{\infty}\frac{\langle n\rangle^n}{n!}\mathcal{L}\{f^{(1)}(\Delta)\}^n$$
$$= \exp\left[-Nx\int_0^{\infty}\mathrm{d}E\left(1-\mathrm{e}^{-sE}\right)\frac{\mathrm{d}\sigma}{\mathrm{d}E}\right].$$

Following Landau [20], we split the integral in the exponent in two parts,

$$Nx\int_0^{\infty}\mathrm{d}E\left(1-\mathrm{e}^{-sE}\right)\frac{\mathrm{d}\sigma}{\mathrm{d}E} = Nx\int_0^{E_1}\mathrm{d}E\left(1-\mathrm{e}^{-sE}\right)\frac{\mathrm{d}\sigma}{\mathrm{d}E} + Nx\int_{E_1}^{\infty}\mathrm{d}E\left(1-\mathrm{e}^{-sE}\right)\frac{\mathrm{d}\sigma}{\mathrm{d}E},$$

where E_1 is chosen to be large compared to the ionisation threshold while at the same time satisfying $sE_1 \ll 1$. For energy transfers exceeding E_1, the differential cross section is assumed to be given by the asymptotic expression for close collisions (2.11); for $E < E_1$, it is not specified.

Using $\exp(-sE) \sim 1 - sE$, we obtain for the first term

$$I_1 = Nx\int_0^{E_1}\mathrm{d}E\frac{\mathrm{d}\sigma}{\mathrm{d}E}\left(1-\mathrm{e}^{-sE}\right) \sim Nxs\int_0^{E_1}\mathrm{d}E\frac{\mathrm{d}\sigma}{\mathrm{d}E}E.$$

We can therefore evaluate I_1 by subtracting the contribution due to energy transfers between E_1 and $E_{\max}$ according to Eq. (2.11) from the total average energy loss $x\mathrm{d}E/\mathrm{d}x = \langle\Delta\rangle$,

$$I_1 \sim s\langle\Delta\rangle - s\xi\left(\ln\frac{E_{\max}}{E_1} - \beta^2\right),$$

where we have introduced the variable

$$\xi = x\frac{2\pi z^2(\alpha\hbar c)^2 NZ}{mc^2\beta^2}.$$

For evaluating the second integral, we approximate $\mathrm{d}\sigma/\mathrm{d}E$ by the Rutherford cross section $\mathrm{d}\sigma_R/\mathrm{d}E \propto 1/E^2$. Because of the rapid convergence of the integral for $sE \gg 1$, we further assume that the upper integration limit can be extended to infinity (instead of truncating $\mathrm{d}\sigma/\mathrm{d}E$ at $E_{\max}$). Integrating by parts and substituting $z = sE$ yields

$$I_2 = \xi \int_{E_1}^{\infty} \mathrm{d}E \frac{1-\mathrm{e}^{-sE}}{E^2} = \xi \underbrace{\frac{1-\mathrm{e}^{-sE_1}}{E_1}}_{\sim s} + \xi s \int_{sE_1}^{\infty} \mathrm{d}z \frac{\mathrm{e}^{-z}}{z} \sim \xi s \left(1 + \int_{sE_1}^{1} \frac{\mathrm{d}z}{z} - C\right),$$

where $C \sim 0.577215665$ is Euler's constant.[3] Combining the two terms I_1 and I_2, one obtains

$$F(s,x) = \exp\left[-\xi s \left(1 - C + \frac{\langle\Delta\rangle}{\xi} - \ln sE_{\max} + \beta^2\right)\right],$$

and, applying the inverse Laplace transform,

$$f(\Delta, x) = \mathcal{L}^{-1}\{F(s,x)\} = \frac{1}{\xi}\phi_L(\lambda), \tag{2.37}$$

where

$$\phi_L(\lambda) = \frac{1}{2\pi\mathrm{i}} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \mathrm{d}u\, \mathrm{e}^{u\ln u + \lambda u}. \tag{2.38}$$

is a universal function of the dimensionless variable

$$\lambda = \frac{\Delta - \langle\Delta\rangle}{\xi} - (1-C) - \beta^2 - \ln\frac{\xi}{E_{\max}}.$$

The maximum of $\phi_L(\lambda)$ is located at $\lambda \sim -0.222782$ and the most probable energy loss is, consequently, given by

$$\Delta_{\mathrm{p}} \sim \langle\Delta\rangle + \xi\left(0.2 + \beta^2 + \ln\kappa\right), \tag{2.39}$$

[3] $$-C = \int_0^1 \mathrm{d}z \frac{\mathrm{e}^{-z}-1}{z} + \int_1^{\infty} \mathrm{d}z \frac{\mathrm{e}^{-z}}{z} = -0.577215665\ldots$$

where $\kappa = \xi/E_{\max}$. The full width at half maximum (FWHM) of the Landau distribution[4] (2.37) is approximately 4.02ξ.

A somewhat unsatisfactory aspect of $\phi_L(\lambda)$ is that its mean is undefined (a consequence of allowing arbitrarily large energy transfers $E > E_{\max}$). This deficiency was overcome by Vavilov [77] who, taking account of the kinematically allowed maximum energy transfer $E_{\max}$ and using the differential cross section (2.11) in I_2, obtained

$$f(\Delta, x) = \frac{1}{\xi}\phi_V(\lambda), \qquad \phi_V(\lambda) = \frac{1}{2\pi \mathrm{i}} \mathrm{e}^{\kappa\left(1+\beta^2 C\right)} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} \exp\left(\psi(u) + \lambda u\right) \mathrm{d}u,$$

where

$$\psi(u) = u \ln \kappa + \left(u + \beta^2 \kappa\right) \left(\int_{u/\kappa}^{\infty} \frac{\mathrm{e}^{-t}}{t} \mathrm{d}t + \ln \frac{u}{\kappa} \right) - \kappa \mathrm{e}^{-u/\kappa}.$$

For small values of κ ($\kappa < 0.01$ [77]) the Vavilov distribution tends to the Landau distribution, while for $\kappa \gg 1$ it approaches a Gaussian distribution with $\sigma^2 = \xi E_{\max}\left(1 - \beta^2/2\right)$ [78]. Algorithms for the numerical evaluation of ϕ_L and ϕ_V and for drawing random numbers from these distributions are discussed e.g. in Refs. [78–81] and are implemented in ROOT [82].

Attempts have been made [83, 84] to improve the Landau-Vavilov method with respect to the treatment of distant collisions by including the second order term in the expansion of $\exp(-sE)$ in I_1. The results are akin to convolving ϕ_L or ϕ_V with a Gaussian distribution (expressions for estimating the standard deviation σ of the Gaussian are reviewed in Ref. [41]).

2.5.4 Examples

Let us first consider track segments for which the projectile suffers on average only tens of collisions. At the minimum of M_0, $\langle n \rangle = 10$ corresponds to a track length $x \sim 4\,\mathrm{mm}$ for argon (at atmospheric pressure, $T = 20\,^{\circ}\mathrm{C}$) and $x \sim 2\,\mu\mathrm{m}$ for silicon (Tables 2.1 and 2.2). As can be seen from Figs. 2.11 and 2.12, the features of the differential cross section $\mathrm{d}\sigma/\mathrm{d}E$ are clearly visible in the straggling functions $f(\Delta, x)$. These spectra cannot be described by a Landau distribution (or variants thereof) and need to be calculated using Monte Carlo simulation or numerical convolution.

[4] In high-energy physics parlance, the term "Landau distribution" is sometimes used for energy loss spectra $f(\Delta, x)$ in general. In this chapter, it refers only to the distribution given by Eq. (2.37).

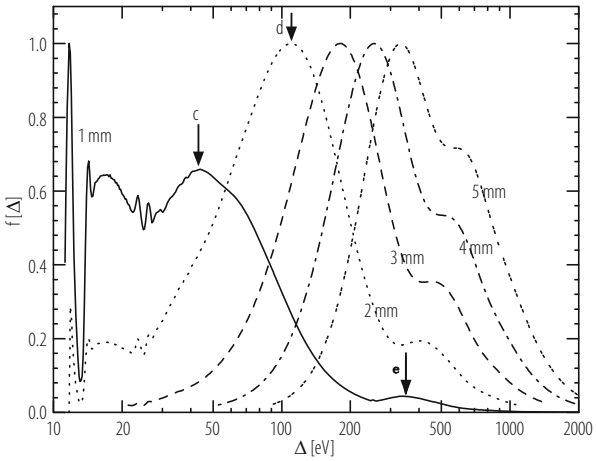


Fig. 2.11 Straggling functions for singly charged particles with $\beta\gamma = 4.48$ traversing segments of length $x = 1 \ldots 5$ mm in Ar. The inverse mean free path M_0 is 30 collisions/cm. The functions are normalised to unity at the most probable value. The broad peak at ∼17 eV is due to single collisions, see Fig. 2.9. For two collisions it broadens and shifts to about 43 eV, marked c, and for $n = 3$ it can be seen at d. It may be noted that the peak at 11.7 eV (if the function is normalised to unit area) is exactly proportional to $\langle n \rangle \exp(-\langle n \rangle)$, as expected from Eq. (2.35). Energy losses to L-shell electrons of Ar (with a binding energy of ∼250 eV) appear at e, for $x = 1$ mm they have an amplitude of 0.04. For $x > 2$ mm, peak c disappears, and peak d becomes the dominant contribution defining the most probable energy loss Δ_p. The buildup for peak e at 440–640 eV is the contribution from L-shell collisions. It appears roughly at 250 eV+Δ_p. The inverse mean free path for collisions with $E > 250$ eV is only 1.7 collisions/cm, thus the amplitude of the peak e is roughly proportional to x. The Bethe mean energy loss is 250 eV/mm

With increasing number of collisions, the detailed features of the differential cross section become "washed out" and the energy loss spectra $f(\Delta, x)$ tend to the Landau shape but are typically broader, as shown in Figs. 2.13 and 2.14. Reasonable agreement with measured energy loss spectra for thin absorbers can often be achieved by fit functions based on the convolution of a Landau/Vavilov distribution and a Gaussian distribution. For a predictive calculation of $f(\Delta, x)$, however, numerical convolution or a Monte Carlo simulation are usually needed.

2.5.5 Methods for Thick Absorbers

In order to compute the energy loss distribution for a layer of material in which the kinetic energies T of the traversing particles change considerably (i.e. by more than 5–10% [86]), we divide the absorber in segments of length x that are sufficiently small such that the straggling function $f(\Delta, x)$ can be calculated using the methods for thin absorbers described above. Let $\phi(y, T)$ be the distribution of

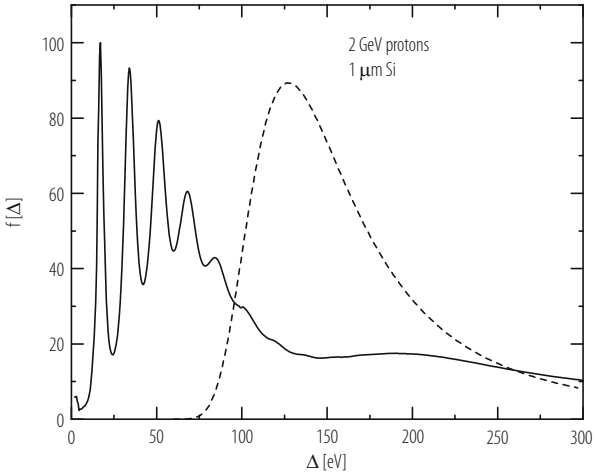


Fig. 2.12 Straggling in 1 μm of Si ($\langle n \rangle = 4$) for particles with $\beta\gamma = 2.1$, compared to the Landau function (dashed line). The Bethe mean energy loss is $\langle \Delta \rangle = 400$ eV. Measured straggling functions of this type are given in Ref. [85]

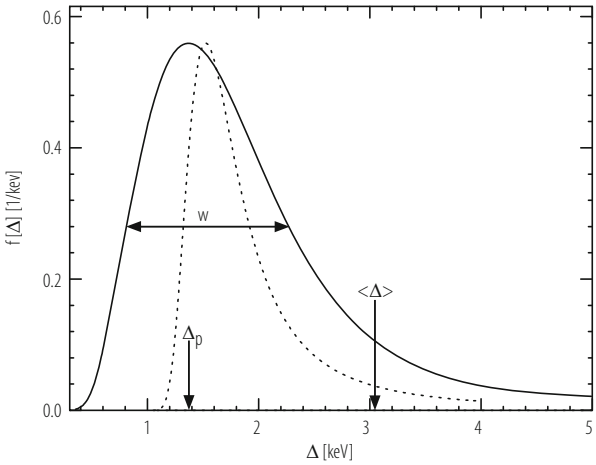


Fig. 2.13 Straggling function $f(\Delta)$ for particles with $\beta\gamma = 3.6$ traversing 1.2 cm of Ar gas ($\langle n \rangle = 36$) calculated using the convolution method (solid line) compared to the Landau distribution (dashed line). Parameters describing $f(\Delta)$ are the most probable energy loss Δ_p, i.e. the position of the maximum of the straggling function, at 1371 eV, and the full width at half maximum (FWHM) $w = 1463$ eV. The Bethe mean energy loss is $\langle \Delta \rangle = 3044$ eV. The peak of the Landau function is at 1530 eV

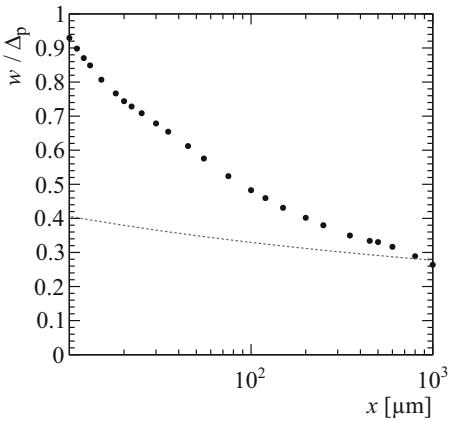


Fig. 2.14 Relative width (full width at half maximum w divided by the most probable value Δ_p) of the straggling spectrum $f(\Delta, x)$ as function of the absorber thickness x, for particles with $\beta\gamma \sim 3.16$ in silicon. The dashed line corresponds to the relative width of the Landau distribution. Circles represent results of a Monte Carlo simulation using the Bethe-Fano differential cross section

kinetic energies at a distance y in the absorber. If $f(\Delta, x)$ is known for all T, the spectrum of kinetic energies at $y + x$ can be calculated using

$$\phi(y + x, T) = \int \phi(y, T + \Delta)\, f(\Delta, x; T + \Delta)\, \mathrm{d}\Delta .$$

Scaling relations, discussed in Ref. [51], can be used to limit the number of thin-absorber distributions $f(\Delta, x; T)$ that need to be tabulated.

In a "condensed history" Monte Carlo simulation [87], the energy loss spectrum is calculated stochastically by sampling the energy loss over a substep x from a suitable thin-absorber distribution (e.g. a Vavilov function), and updating the kinetic energy T of the projectile after each substep.

2.6 Energy Deposition

Leaving the emission of Cherenkov radiation and other collective effects aside, charged-particle collisions with electrons in matter result in the promotion of one of the electrons in the target medium to a bound excited state or to the continuum. Both effects (excitation and ionisation) can be exploited for particle detection purposes. In scintillators, discussed in Chap. 3 of this book, part of the energy transferred to excitations is converted to light. Detectors based on ionisation measurement in gases and semiconductors are discussed in Chaps. 4 and 5. In the following we briefly review the main mechanisms determining the number of electron-ion pairs (in gases) or electron-hole pairs (in semiconductors) produced in the course of an ionising collision, along with their spatial distribution.

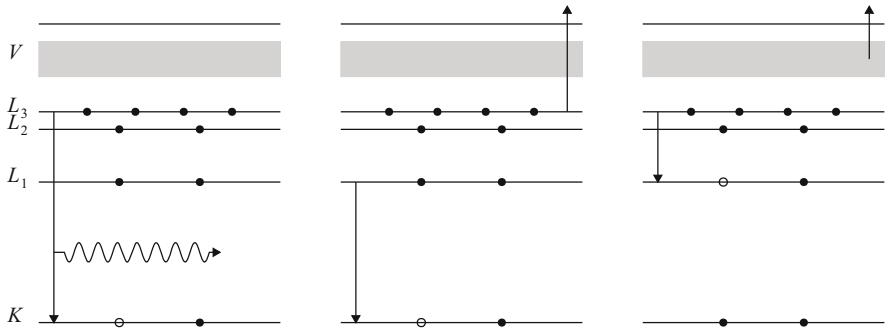


Fig. 2.15 After the ejection of an inner-shell electron, the resulting vacancy is filled by an electron from a higher shell. The energy released in the transition can either be carried away by a fluorescence photon (left) or be transferred to an electron in a higher shell (Auger process, middle). Coster-Kronig transitions (right) are Auger processes in which the initial vacancy is filled by an electron from the same shell

2.6.1 Atomic Relaxation

If a charged-particle collision (or a photoabsorption interaction) ejects an inner-shell electron from an atom, the resulting vacancy will subsequently be filled by an electron from a higher shell, giving rise to a relaxation chain which can proceed either radiatively, i.e. by emission of a fluorescence photon, or radiation-less (Auger effect). The two processes are illustrated schematically in Fig. 2.15. Fluorescence photons can in turn ionise another atom in the medium or, with a probability depending on the geometry of the device, escape from the detector. The fluorescence yield, i.e. the probability for a vacancy to be filled radiatively, increases with the atomic number Z: in silicon, for example, the average fluorescence yield is $\sim$5%, compared to $\sim$54% in germanium [88]. Compilations of fluorescence yields can be found in Refs. [88–91]. Tabulations of transition probabilities are available in the EADL database [92, 93].

2.6.2 Ionisation Statistics

The "primary" ionisation electron knocked out in a collision (and also the Auger electrons) may have kinetic energies exceeding the ionisation threshold of the medium and thus undergo further ionising collisions along their path. Electrons with a kinetic energy T that is large compared to the ionisation threshold are referred to as "delta" electrons; their energy distribution follows approximately the close-collision differential cross section, given by Eq. (2.11) for spin-zero particles. The number of electrons n_e produced in the energy degradation cascade of a delta electron with initial kinetic energy T is subject to fluctuations. The mean and variance of the

distribution of n_e are described by the average energy W required to produce an electron-ion (electron-hole) pair,

$$\langle n_e \rangle = \frac{T}{W}, \tag{2.40}$$

and the Fano factor F [94],

$$\sigma^2 = \langle (n - \langle n \rangle)^2 \rangle = F \langle n_e \rangle, \tag{2.41}$$

respectively. Both W and F are largely determined by the relative importance of ionising and non-ionising inelastic collisions, the latter including e.g. excitations or phonon scattering. If the cross sections for these processes are known, the distribution of n_e can be calculated using detailed Monte Carlo simulations. An example is the MAGBOLTZ program [95, 96], which includes the relevant cross sections for many commonly used detection gases. Inelastic cross sections of delta electrons in solids can be calculated based on the dielectric formalism discussed in Sect. 2.3.1 (in its non-relativistic version), often making using of optical data and a suitable model of the q-dependence of $\mathrm{Im}\,(-1/\varepsilon\,(q, E))$ as, for instance, in the Penn algorithm described in Ref. [97].

Measurements of W for electrons in gases as a function of the electron's initial kinetic energy are reported in Refs. [98–100, 102, 103]. As can be seen from Fig. 2.16, which shows measurements and calculations for CO_2, W increases towards low kinetic energies, while in the keV range and above it depends only weakly on T. For most gases and semiconductors typically used as sensitive media in particle detectors, the asymptotic (high-energy) W values are fairly well established. A compilation of recommended average W values, based on experimental data until 1978, is given in ICRU report 31 [101]. Critical reviews of W values and Fano factors including also more recent data can be found in Ref. [104]

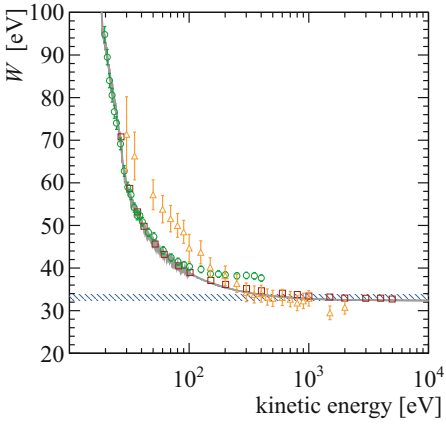


Fig. 2.16 W value for electrons in CO_2 as a function of the electron's initial kinetic energy according to measurements by Combecher [98] (circles), Smith and Booz [99] (triangles), and Waibel and Grosswendt [100] (squares). The grey band represents results of a Monte Carlo calculation using the cross sections implemented in MAGBOLTZ [96]. The hatched band corresponds to the high-energy value recommended in Ref. [101]

Table 2.3 Asymptotic W values and Fano factors for different gases and for solid silicon (at $T = 300\,\mathrm{K}$)

	W [eV]	F
Ne	35.4 ± 0.9 [101]	0.13–0.17 [104]
Ar	26.4 ± 0.5 [101]	0.15–0.17 [104]
Kr	24.4 ± 0.3 [101]	0.17–0.21 [104]
Xe	22.1 ± 0.1 [101]	0.124–0.24 [104]
CO_2	33.0 ± 0.7 [101]	0.32 [104]
CH_4	27.3 ± 0.3 [101]	0.22–0.26 [104]
iC_4H_{10}	23.4 ± 0.4 [101]	0.261 [106]
CF_4	34.3 [107]	
Si	3.67 ± 0.02 [108]	<0.1 [104]

Except for CF_4, the values shown are for measurements using electrons

and, with emphasis on noble gases, in Ref. [105]. Parameters for silicon and some commonly used gases are listed in Table 2.3.

Analogously to Eqs. (2.40) and (2.41) one can define W values and Fano factors characterising the distribution of the number of electrons produced by a heavy charged particle (provided that it is stopped completely in the medium) or by the absorption of a photon. The asymptotic W values for electrons and photons at high energies are in general very similar.

In gas mixtures without excitation transfers, the W value and Fano factor are, to a good approximation, given by the values in the pure gases, weighted by their respective concentrations. In mixtures where one of the components has excited states with energies exceeding the ionisation threshold of another component, excitation transfer can lead to a significant reduction of W and F with respect to the pure gases ("Jesse effect" [109]). Results for a number of binary gas mixtures from measurements with α particles can be found in Ref. [110].

2.6.3 Range

The spatial distribution of secondary ionisations produced by a delta electron can be characterised in terms of the electron range, i.e. the typical path length travelled by an electron before its energy falls below the ionisation threshold. In the literature, a number of different definitions of "range" exist, two of which—the fractional ionisation range R_x and the practical range R_p—are illustrated in Fig. 2.17. If the cross sections (including those for elastic scattering) are known, the range of delta electrons and, more generally, the ionisation pattern produced by a charged-particle collision, can be calculated using Monte Carlo techniques. As an example, Fig. 2.18 shows measurements of the 95% range in CH_4 as a function of the primary electron energy [102], together with calculated values based on the cross sections implemented in MAGBOLTZ.

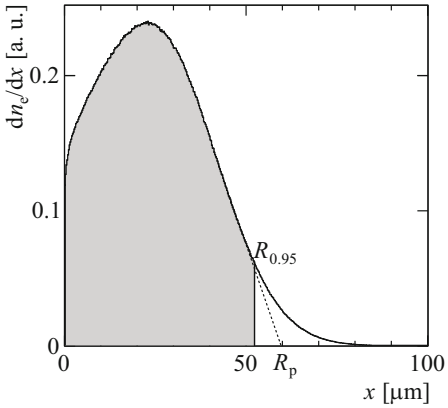


Fig. 2.17 Distribution of the coordinates (projected on the electron's initial direction) of ionising collisions by a $T = 1$ keV electron and its secondaries in methane (at atmospheric pressure, $T = 20\,^{\circ}\mathrm{C}$), calculated using the cross sections implemented in MAGBOLTZ. The fractional ionisation range R_x is defined as the projected distance along the electron's initial direction within which the fraction x of the total ionisation is produced [102]. The practical range R_p is determined by linear extrapolation from the region of steepest descent to the horizontal axis

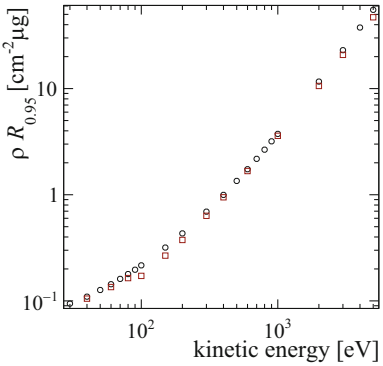


Fig. 2.18 Measurements [102] (squares) and MAGBOLTZ calculations (circles) of the 95% fractional ionisation range of electrons in methane (at atmospheric pressure)

In the absence of a detailed calculation, the semi-empirical formula by Kobetich and Katz [111, 112] can be used to estimate the practical range,

$$\rho R_p\,(T) = AT\left(1 - \frac{B}{1 + CT}\right),$$

where the parameters A, B, C are given by [112]

$$A = \left(0.81Z^{-0.38} + 0.18\right) \times 10^{-3}\,\mathrm{g\,cm^{-2}\,keV^{-1}},$$
$$B = 0.21Z^{-0.055} + 0.78,$$
$$C = \left(1.1Z^{0.29} + 0.21\right) \times 10^{-3}\,\mathrm{keV^{-1}}.$$

References

1. U. Fano, Annu. Rev. Nucl. Sci. **13**, 1 (1963).
2. M. Inokuti, Rev. Mod. Phys. **43**, 297 (1971).
3. M. Inokuti, Rev. Mod. Phys. **50**, 23 (1978).
4. H. Bichsel, Passage of Charged Particles Through Matter, in *Amer. Instrum. Phys. Handbook*, edited by D. E. Gray, McGraw Hill, 1972.
5. R. J. Gould, Physica **62**, 555 (1972).
6. W. W. M. Allison and J. H. Cobb, Annu. Rev. Nucl. Part. Sci. **30**, 253 (1980).
7. S. P. Ahlen, Rev. Mod. Phys. **52**, 121 (1980).
8. B. Rossi, *High-Energy Particles*, Prentice-Hall, New York, NY, 1952.
9. R. D. Evans, *The Atomic Nucleus*, McGraw Hill, 1967.
10. W. Heitler, *The Quantum Theory of Radiation, second edition*, Oxford University Press, 1949.
11. P. Sigmund, *Particle Penetration and Radiation Effects*, Springer, 2006.
12. N. J. Carron, *An Introduction to the Passage of Energetic Particles through Matter*, Taylor and Francis, 2007.
13. B. McParland, *Medical radiation dosimetry: Theory of charged particle collision energy loss*, Springer, 2014.
14. N. Bohr, Philos. Mag. **25**, 10 (1913).
15. N. Bohr, Philos. Mag. **30**, 581 (1915).
16. H. Bethe, Ann. Phys. **5**, 325 (1930).
17. H. Bethe, Z. Phys. **76**, 293 (1932).
18. E. Fermi, Z. Phys. **29**, 3157 (1924).
19. E. Fermi, Phys. Rev. **57**, 485 (1940).
20. L. D. Landau, J. Phys. USSR **8**, 201 (1944).
21. U. Fano, L. V. Spencer, and M. J. Berger, Penetration and Diffusion of X Rays, in *Handbuch der Physik*, edited by S. Flügge, volume 38/2, Springer, 1959.
22. W. Rösch, E. Tochilin, and F. H. Attix, editors, *Radiation Dosimetry; 2nd ed.*, Academic Press, 1968.
23. F. Salvat and J. M. Fernandez-Varea, Metrologia **46**, S112 (2009).
24. M. J. Berger et al., XCOM: Photon Cross Sections Database, https://www.nist.gov/pml/xcom-photon-cross-sections-database.
25. J. Berkowitz, *Atomic and molecular photo absorption*, Academic Press, 2002.
26. N. Sakamoto et al., Oscillator Strength Spectra and Related Quantities of 9 Atoms and 23 Molecules Over the Entire Energy Region, NIFS-DATA-109, National Institute for Fusion Science, 2010, available online at http://www.nifs.ac.jp/report/nifsdata.html.
27. B. L. Henke, E. M. Gullikson, and J. C. Davis, Atomic Data and Nucl. Data Tables **54**, 181 (1993), available online at http://henke.lbl.gov/optical_constants/.
28. J.-J. Yeh and I. Lindau, Atomic Data and Nucl. Data Tables **32**, 1 (1985).

29. M. B. Trzhaskovskaya, V. I. Nefedov, and V. G. Yarzhemsky, Atomic Data and Nucl. Data Tables **77**, 97 (2001).
30. M. B. Trzhaskovskaya, V. I. Nefedov, and V. G. Yarzhemsky, Atomic Data and Nucl. Data Tables **82**, 257 (2002).
31. E. D. Palik and E. J. Prucha, *Handbook of optical constants of solids*, Academic Press, 1998.
32. S. Adachi, *Optical constants of crystalline and amorphous semiconductors*, Kluwer Academic Publishers, 1999.
33. O. Klein and Y. Nishina, Z. Physik **52**, 853 (1929).
34. J. H. Hubbell, Photon Cross Sections, Attenuation Coefficients, and Energy Absorption Coefficients from 10 keV to 100 GeV, NSRDS-NBS 29, National Bureau of Standards, 1969, available online at https://www.nist.gov/srd/national-standard-reference-data-series.
35. P. M. Bergstrom and R. H. Pratt, Radiation Physics and Chemistry **50**, 3 (1997).
36. J. H. Hubbell, H. A. Gimm, and I. Øverbø, J. Phys. Chem. Ref. Data **9**, 1023 (1980).
37. D. Bote and F. Salvat, Phys. Rev. A **77**, 042701 (2008).
38. L. D. Landau, E. M. Lifshitz, and L. P. Pitaevskii, *Electrodynamics of Continuous Media*, Butterworth, 1984.
39. J. Lindhard, Mat. Fys. Medd. Dan. Vid. Selsk. **28**, 1 (1954).
40. J. Lindhard and A. Winther, Mat. Fys. Medd. Dan. Vid. Selsk. **34**, 1 (1964).
41. H. Bichsel, Rev. Mod. Phys. **60**, 663 (1988).
42. H. Bichsel, Phys. Rev. A **65**, 052709 (2002).
43. H. Bichsel, Phys. Rev. A **46**, 5761 (1992).
44. H. Bichsel, K. M. Hanson, and M. E. Schillaci, Phys. Med. Biol. **27**, 959 (1982).
45. M. C. Walske, Phys. Rev. **88**, 1283 (1952).
46. M. C. Walske, Phys. Rev. **101**, 1940 (1956).
47. V. A. Chechin, L. P. Kotenko, G. I. Merson, and V. C. Yermilova, Nucl. Instr. Meth. **98**, 577 (1972).
48. F. Lapique and F. Piuz, Nucl. Instr. Meth. **175**, 297 (1980).
49. I. B. Smirnov, Nucl. Instr. Meth. A **554**, 474 (2005).
50. J. Apostolakis et al., Nucl. Instr. Meth. A **453**, 597 (2000).
51. H. Bichsel, Nucl. Instrum. Meth. A **562**, 154 (2006).
52. F. Rieke and W. Prepejchal, Phys. Rev. A **6**, 1507 (1972).
53. R. P. Saxon, Phys. Rev. A **8**, 839 (1973).
54. M. Inokuti, R. P. Saxon, and J. L. Dehmer, Int. J. Radiat. Phys. Chem. **7**, 109 (1975).
55. B. Sitar, G. I. Merson, V. A. Chechin, and Y. A. Budagov, *Ionization Measurements in High Energy Physics*, Springer, 1993.
56. G. W. McClure, Phys. Rev. **90**, 796 (1953).
57. G. Malamud, A. Breskin, R. Chechik, and A. Pansky, J. Appl. Phys. **74**, 3645 (1993).
58. H. A. Bethe and R. W. Jackiw, *Intermediate quantum mechanics*, Benjamin, 1986.
59. R. M. Sternheimer, M. J. Berger, and S. M. Seltzer, Atomic Data and Nucl. Data Tables **30**, 261 (1984).
60. *Stopping powers and ranges for protons and alpha particles*, International Commission on Radiation Units and Measurements, Washington, DC, 1993, ICRU Report 49.
61. M. J. Berger, J. S. Coursey, M. A. Zucker, and J. Chang, ESTAR, PSTAR, and ASTAR: Computer Programs for Calculating Stopping-Power and Range Tables for Electrons, Protons, and Helium Ions, http://physics.nist.gov/Star.
62. D. E. Groom, N. V. Mokhov, and S. I. Striganoff, Atomic Data and Nucl. Data Tables **78**, 183 (2001).
63. W. H. Barkas, J. N. Dyer, and H. H. Heckman, Phys. Rev. Lett. **11**, 26 (1963).
64. H. H. Andersen, H. Simonsen, and H. Sørensen, Nuclear Physics A **125**, 171 (1969).
65. H. Bichsel, Phys. Rev. A **41**, 3642 (1990).
66. E. A. Uehling, Annu. Rev. Nucl. Sci. **4**, 315 (1954).
67. *Stopping powers for electrons and positrons*, International Commission on Radiation Units and Measurements, Washington, DC, 1979, ICRU Report 37.
68. H. Bethe and W. Heitler, Proc. Roy. Soc. A **146**, 83 (1934).

69. H. W. Koch and J. W. Motz, Rev. Mod. Phys. **31**, 920 (1959).
70. M. Tanabashi et al., Phys. Rev. D **98**, 030001 (2018).
71. D. E. Groom, Atomic and Nuclear Properties of Materials, http://pdg.lbl.gov/AtomicNuclearProperties.
72. E. J. Williams, Proc. Roy. Soc. A **125**, 420 (1929).
73. H. Bichsel and R. P. Saxon, Phys. Rev. A **11**, 1286 (1975).
74. F. Salvat, J. M. Fernández-Varea, and J. Sempau, PENELOPE-2014: A Code System for Monte Carlo Simulation of Electron and Photon Transport.
75. A. M. Kellerer, *Mikrodosimetrie*, Strahlenbiologisches Institut der Universität München, 1968, G.S.F. Bericht B-1.
76. A. M. Kellerer, Fundamentals of Microdosimetry, in *The Dosimetry of Ionizing Radiation*, edited by K. R. Kase, B. E. Bjängard, and F. H. Attix, Academic Press, 1985.
77. P. V. Vavilov, Sov. Phys. JETP **5**, 749 (1957).
78. M. J. Berger and S. M. Seltzer, *Studies in Penetration of Charged Particles in Matter*, The National Academies Press, Washington, DC, 1964.
79. W. Boersch-Supan, J. Res. Nat. Bur. Stand. **65B**, 245 (1961).
80. B. Schorr, Computer Phys. Comm. **7**, 215 (1974).
81. A. Rotondi and P. Montagna, Nucl. Instr. Meth. B **47**, 215 (1990).
82. R. Brun et al., Nucl. Instrum. Meth. A **389**, 81 (1997), http://root.cern.ch.
83. O. Blunck and S. Leisegang, Z. Phys. **128**, 500 (1950).
84. P. Shulek et al., Sov. J. Nucl. Phys. **4**, 400 (1967).
85. J. P. Perez, J. Sevely, and B. Jouffrey, Phys. Rev. A **16**, 1061 (1977).
86. C. Tschalär, Nucl. Instrum. Meth. **61**, 141 (1968).
87. M. J. Berger, Meth. Comput. Phys. **1**, 135 (1963).
88. M. O. Krause, J. Phys. Chem. Ref. Data **8**, 307 (1979).
89. W. Bambynek et al., Rev. Mod. Phys. **44**, 716 (1972).
90. J. H. Hubbell et al., J. Phys. Chem. Ref. Data **23**, 339 (1994).
91. J. H. Hubbell et al., J. Phys. Chem. Ref. Data **33**, 621 (2004).
92. S. T. Perkins et al., Tables and Graphs of Atomic Subshell and Relaxation Data Derived from the LLNL Evaluated Atomic Data Library (EADL), $Z = 1 - 100$, UCRL-50400, Lawrence Livermore National Laboratory, Livermore, CA, 1991.
93. D. E. Cullen, EPICS2017. Electron and Photon Interaction Cross Sections, https://www-nds.iaea.org/epics/.
94. U. Fano, Phys. Rev. **72**, 26 (1947).
95. S. F. Biagi, Nucl. Instr. Meth. A **421**, 234 (1999).
96. S. F. Biagi, Magboltz 8, http://magboltz.web.cern.ch/magboltz.
97. D. R. Penn, Phys. Rev. B **35**, 482 (1987).
98. D. Combecher, Radiation Research **84**, 189 (1980).
99. B. G. R. Smith and J. Booz, Experimental results on W-values and transmission of low energy electrons in gases, in *Sixth Symposium on Microdosimetry*, page 759, 1978.
100. E. Waibel and B. Grosswendt, Nucl. Instr. Meth. B **53**, 239 (1991).
101. *Average energy required to produce an ion pair*, International Commission on Radiation Units and Measurements, Washington, DC, 1979, ICRU Report 31.
102. E. Waibel and B. Grosswendt, Nucl. Instr. Meth. **211**, 487 (1983).
103. E. Waibel and B. Grosswendt, Study of W-values, practical ranges, and energy dissipation profiles of low-energy electrons in N_2, in *Eighth Symposium on Microdosimetry*, page 301, 1983.
104. Atomic and Molecular Data for Radiotherapy and Radiation Research, IAEA TECDOC 799, 1995.
105. I. Krajcar Bronić, Hoshasen: ionizing radiation **24**, 101 (1998).
106. D. Srdoč, B. Obelić, and I. Krajcar Bronić, J. Phys. B: At. Mol. Phys. **20**, 4473 (1987).
107. G. F. Reinking, L. G. Christophorou, and S. R. Hunter, J. Appl. Phys. **60**, 499 (1986).
108. R. H. Pehl, F. S. Goulding, D. A. Landis, and M. Lenzlinger, Nucl. Instr. Meth. **59**, 45 (1968).
109. W. P. Jesse and J. Sadauskis, Phys. Rev. **88**, 417 (1952).

110. T. E. Bortner, G. S. Hurst, M. Edmundson, and J. E. Parks, Alpha particle ionization of argon mixtures – further study of the role of excited states, ORNL-3422, Oak Ridge National Laboratory, 1963.
111. E. J. Kobetich and R. Katz, Phys. Rev. **170**, 391 (1968).
112. E. J. Kobetich and R. Katz, Nucl. Instr. Meth. **71**, 226 (1969).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 3
Scintillation Detectors for Charged Particles and Photons

P. Lecoq

3.1 Basic Detector Principles and Scintillator Requirements

3.1.1 Interaction of Ionizing Radiation with Scintillator Material

As any radiation detector, a scintillator is an absorbing material, which has the additional property to convert into light a fraction of the energy deposited by ionizing radiation. Charged and neutral particles interact with the scintillator material through the well-known mechanisms of radiation interactions in matter described by many authors [1, 2]. Charged particles continuously interact with the electrons of the scintillator medium through Coulomb interactions, resulting in atomic excitation or ionization. Neutral particles will first have to undergo a direct interaction with the nucleus producing recoil protons or spallation fragments, which will then transfer their energy to the medium in the same way as primary charged particles.

The rate of energy loss $(-\mathrm{d}E/\mathrm{d}x)$ for charged particles is strongly energy dependant. It is well described by the Bethe-Bloch formula (see Chap. 2) for incoming particles in the MeV-GeV range, with atomic shell corrections at lower energy and radiative loss corrections at higher energy. For heavy materials currently used as scintillators with a density of 6–8 g/cm^3, it is typically of the order of 10 MeV/cm for a minimum ionizing particle but it can be a factor up to 100 more at very low or very high energy (radiative losses).

P. Lecoq (✉)
CERN, Geneva, Switzerland
e-mail: Paul.Lecoq@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_3

In the case of X- or γ- rays, the three fundamental mechanisms of electromagnetic interaction are [3]:

- Photo-absorption
- Compton scattering
- Electron-positron pair production

The dominant process at low energy (up to a few hundred keV for heavy materials) is the photoelectric absorption. The interacting photon transfers its energy to an electron from one of the electron shells of the absorber atom (usually from a deep shell). The resulting photoelectron is ejected with a kinetic energy corresponding to the incident photon energy minus the binding energy of the electron on its shell. This is followed by a rapid reorganization of the electron cloud to fill the electron vacancy, which results in the emission of characteristic X-Rays or Auger electrons. The photoelectric process has the highest probability when the incident photon has an energy comparable to the kinetic energy of the electron on its shell. This is the origin of the typical peaks observed in the cross-section curve corresponding to resonances for the different electron shells (Fig. 3.1). The general trend of this cross-section is a rapid decrease with energy and a strong dependence on the atomic number Z of the absorber explaining the preponderance of high-Z materials for X- or γ-rays detection and shielding:

$$\sigma_{\text{ph}} \propto \frac{Z^5}{E_\gamma^{7/2}} \tag{3.1}$$

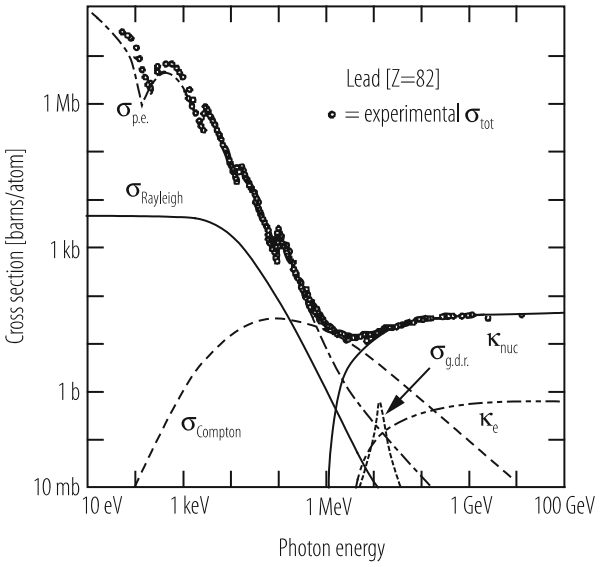


Fig. 3.1 Energy dependence of photon total cross sections in Lead (from Particle Data Group)

At energies above a few hundred keV, Compton scattering becomes predominant. In this case, the incident photon transfers only part of its initial energy E_γ to an electron of the atomic shells and is scattered at an angle θ with respect to its original direction. The recoil electron is then rapidly absorbed by the scintillator and releases an energy according to the formula:

$$E_\mathrm{e} = E_\gamma - E'_\gamma - E_\mathrm{ebinding} \tag{3.2}$$

where E'_γ is the energy of the scattered photon given by (with m_0 the rest mass of the electron):

$$E'_\gamma = \frac{E_\gamma}{1 + \frac{E_\gamma}{m_0 c^2}(1 - \cos\theta)} \tag{3.3}$$

The energy released in the scintillator by the recoil electron is distributed on a continuum between zero and a maximum up to $E_\gamma - m_0c^2/2 = E_\gamma - 256$ keV (for gamma energy large compared to the rest mass of the electron).

The probability of Compton scattering is related to the electron density in the medium and increases linearly with the atomic number of the absorber, favouring therefore high Z materials.

Above a threshold of 1.02 MeV (twice the rest mass of the electron), the mechanism of e^+e^- pair production can take place, predominantly in the electric field of the nuclei, and to a lesser extent in the electric field of the electron cloud (respectively κ_nuc and κ_e in Fig. 3.1). Similarly to photo-absorption and Compton scattering this process has a higher probability for high Z materials as the cross section is approximately given by the formula [4]:

$$\sigma_\mathrm{pair} \propto Z^2 \ln\left(2E_\gamma\right) \tag{3.4}$$

Below the threshold of electron-positron pair production electrons will continue to loose energy mainly through Coulomb scattering.

In the case of an ordered material like a crystal another mechanism takes place at this stage. In the process of energy degradation the electrons in the keV range start to couple with the electrons of the atoms of the lattice and excite the electrons from the occupied valence or core bands to different levels in the conduction band. Each of these interactions results in an electron-hole pair formation. If the energy of the electron is high enough to reach the ionization threshold free carriers are produced, which will move randomly in the crystal until they are trapped by a defect or recombine on a luminescent centre. In the case the ionization threshold is not reached the electron and hole release part of their energy by coupling to the lattice vibration modes until they reach the top of the valence band for the hole and the bottom of the conduction band for the electron. They can also be bound and form an exciton whose energy is in general slightly smaller than the bandgap between the valence and the conduction bands. At this stage the probability is maximum

for their relaxation on luminescent centres through an energy or a charge transfer mechanism.

For a material to be a scintillator it must contain luminescent centres. They are either extrinsic, generally doping ions, or intrinsic i.e. molecular systems of the lattice or of defects of the lattice, which possess a radiative transition between an excited and a lower energy state. Moreover, the energy levels involved in the radiative transition must be smaller than the forbidden energy bandgap, in order to avoid re-absorption of the emitted light or photo-ionization of the centre.

In a way, a scintillator can be considered as a wavelength shifter. It converts the energy (or wavelength) of an incident particle or energetic photon (UV, X-ray or gamma-ray) into a number of photons of much lower energy (or longer wavelength) in the visible or near visible range, which can be detected by photomultipliers, photodiodes or avalanche photodiodes.

3.1.2 Important Scintillator Properties

Scintillators are among the most popular ionizing radiation detectors.

There are two main classes of scintillators: inorganic and organic. For the inorganic systems (generally ionic crystals), scintillation arises from thermalized electrons and holes, moved to the bottom of the conduction band or the top of the valence band respectively, by scattering from the initially produced fast charge carriers. For the organic systems, scintillation arises upon transition between an excited molecular level and the corresponding electronic ground state. Inorganic scintillators are generally brighter but with a slower decay time than organic ones. However no "ideal" material exists and the choice of a scintillator depends on the application, as it is generally driven by a trade-off between a number of physico-chemical and optical parameters such as density, scintillation properties and radiation hardness. The production and processing cost is also an important issue taking into consideration the very large volumes required for some applications.

3.1.2.1 Physico-chemical Properties

Physico-chemical properties are related to the material composition, structure and density, as well as to its chemical stability when exposed to different environmental conditions: air, humidity, ionizing radiation.

Frequently the density and hence the compactness of the detector is essential in order to reduce the detector volume and cost. This is achieved by using high stopping power and therefore high density materials. This reduces the size of the shower for high energy γ's and electrons as well as the range of Compton scattered photons for lower energy γ-rays. A dense material also reduces the lateral spread of the shower, which is particularly important for the majority of High Energy Physics detectors.

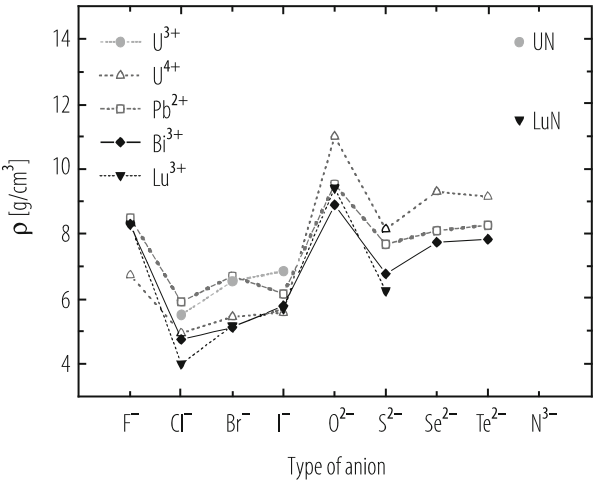


Fig. 3.2 Density for various binary compounds as a function of the binding anion (courtesy P. Derenbos, from ref. [5])

Crystals with a density higher than 8 g/cm^3 are currently available, such as Lead Tungstate (PWO: 8.28 g/cm^3) or Lutetium Aluminium Perovskite (LuAP: 8.34 g/cm^3). Materials of even higher density in the range of 10 g/cm^3 are being identified and studied, such as: Lutetium Oxyde: Lu_2O_3, Lutetium Hafnate: $Lu_4Hf_3O_{12}$, Lutetium Tantalate: Lu_3TaO_7, Lutetium Lead Tantalate: $LuPb_2TaO_6$, Thorium Oxyde: ThO_2. Scintillators are wide bandgap ionic materials and high density implies the choice of anions and cations of high atomic number A (and therefore high Z), as well as small ionic radius to increase the ionic density in the crystal lattice. From this point of view, oxides are generally denser than iodides because of the much smaller ionic radius of the oxygen compared to the iodine ion and in spite of its lighter weight. Similarly, the oxidation potential of the anion is important as it allows reducing the number of anions (generally light) needed to compensate for the positive charge of the much heavier cation. For this reason oxygen is a better ligand than the slightly heavier fluorine ion because of its higher oxidation state (2 or 3 instead of 1). Figure 3.2 illustrates this effect for a number of binary compounds as a function of the anion type.

High Z materials are also preferred for low and medium energy spectroscopy because of the strong dependence of the photoelectric cross-section on Z (see Sect. 3.1.1). High density is also required at high energy to achieve a small radiation length X_0 (mean distance over which an electron loses 1/e of its energy) given as a function of the density ρ, atomic mass A and atomic number Z by:

$$X_0 = \frac{A}{\rho} \frac{716.4 \text{gcm}^{-2}}{Z\,(Z+1)\ln(287/Z)} \tag{3.5}$$

However, contrary to a common assumption, the optimum conditions are not necessarily achieved with the highest Z ions, because in addition to a small X_0, the density ρ should be high. This reduces the lateral shower size given by the Moliere radius:

$$R_{\mathrm{M}} \approx X_0 \cdot (Z + 1.2)\,/37.74 \sim 1/\rho \tag{3.6}$$

The stability of the physico-chemical parameters is also important for the detector design. Scintillation crystals are very stable materials, at least in the bulk, if grown under conditions allowing a good structural quality. This provides a high degree of internal symmetry in the material together with high energetic stability. However, the charge unbalance on the surface can be at the origin of different problems, such as a concentration of impurities or crystallographic defects. As a result, the material can interact with its environment and locally change its properties. The majority of halide crystals have the anions weakly bound to the cations at the surface. They are therefore easily replaced by OH^- radicals from the atmosphere, which have strong optical absorption bands in the visible spectrum. This causes a progressive brownish discoloration of the crystal surface, a well know feature of hygroscopic materials. Encapsulating the crystal in an inert atmosphere avoids this effect.

3.1.2.2 Optical Properties

Inorganic scintillators usually show wide emission bands because of multi-site emission centres differently distorted by the crystal field, as well as by temperature broadening of the optical transitions through vibronic coupling of the emission centres with the crystal lattice. These emission bands are situated in the optical window of the scintillator and produce light in the visible, near infrared or near ultraviolet part of the spectrum. One of the objectives of scintillator development is to design scintillators with emissions peaks matching the maximum quantum efficiency of photodetectors, typically 250–500 nm for photomultipliers and 450–900 nm for solid state photodetectors (pin diodes and avalanche photodiodes).

Light yield (LY) is an essential parameter for a scintillator as it directly influences the energy resolution at low or medium energy through the photostatistic term proportional to $(LY)^{-1/2}$ and the timing resolution proportional to $(\tau_{\mathrm{sc}}/LY)^{-1/2}$, with τ_{sc} being the scintillation decay time. The scintillation mechanism is a multi-step process, which will be described in detail in Sect. 3.2. The overall scintillation yield is determined by the product of efficiencies for all these steps. The dominant factor, which sets the fundamental limit on the light output of a given scintillator, is the number n_{eh} of thermalized electron-hole pairs (active for scintillation) produced in the ionization track of the incoming particle:

$$n_{\mathrm{eh}} = \frac{E_\alpha}{\beta \cdot E_{\mathrm{g}}} \tag{3.7}$$

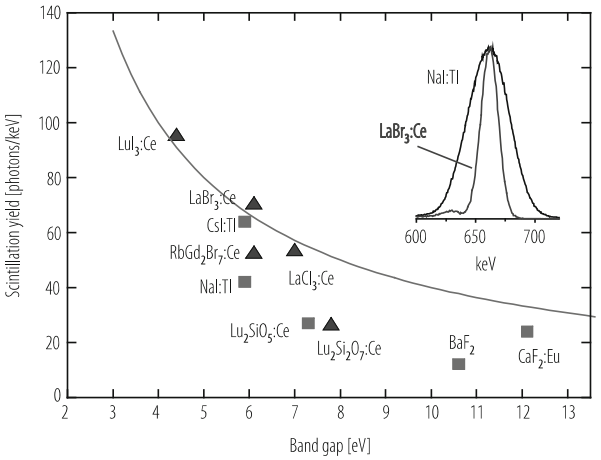


Fig. 3.3 Photon yield/keV of several scintillators as a function of the width of the forbidden band (courtesy P. Dorenbos)

where $\beta \cdot E_g$ is the mean energy necessary for the formation of one thermalized electron-hole pair in a medium with a forbidden zone of width E_g and E_α is the absorbed energy. For ionic crystals, the factor β is usually close to 2.3 and takes into account the energy loss through coupling with lattice phonons during the thermalization process [5]. As shown on Fig. 3.3 low bandgap materials have higher scintillation yields, although such materials are potentially more subject to trap induced quenching, re-absorption phenomena and photo-ionization of the luminescence centre. The ultimate light yield obtained for a material having a bandgap of 3 eV and an emission wavelength of about 600 nm is in the range of 140 photons/keV. The observed signal in photoelectrons/MeV is much smaller, due to losses in the light transport to the photodetector and the quantum efficiency of the photodetector.

The scintillation kinetics is another important consideration as a fast response and low dead time is frequently required for high detection rates. It is related to the rate of decrease of the population of the excited luminescent centres. For a simple process, with only one radiating centre and no interaction between luminescent centres and traps, the decay is exponential and characterized by a time constant τ_{sc}, the time after which the population has decreased by a factor e. For two independent radiating centres the same description with two exponentials holds. Real cases are however very often more complex, involving energy transfer between centres and quenching mechanisms, and the resulting light emission is strongly non-exponential. It is nevertheless common practice to describe this complex emission curve by a sum of exponentials with different time constants. This has in most of the cases no physical justification but simplifies the calculations. If we assume a very fast transfer of the electrons and holes to the luminescent centres the ultimate

limit for the scintillation decay time is given by the transition probability between its excited and ground states:

$$\Gamma = \frac{1}{\tau_{sc}} \propto \frac{n}{\lambda_{em}^3}\left(\frac{n^2+2}{3}\right)^2 \sum_f |\langle f|\ \mu\ |i\rangle|^2 \tag{3.8}$$

where n is the refractive index of the crystal, λ_{em} the emission wavelength of the transition, f and i the wave functions of the final and initial states respectively. The strength of the dipole operator μ connecting the initial and final state determines the decay time of the transition. This matrix element can only be sufficiently large for a transition between two states with different parity (parity allowed transition). This is in particular the case for the 5d to 4f transition in commonly used activators like Ce^{3+}, Pr^{3+}, Nd^{3+} and Eu^{3+}. Forbidden transitions are generally characterized by long decay times, unless a competitive non-radiative relaxation channel exists, which will contribute to the decrease of the population of excited states:

$$\frac{dn_e}{dt} = -\frac{n_e}{\tau} - \alpha n_e e^{-\frac{E}{kT}} \tag{3.9}$$

Here n_e represents the electronic density of the excited state, which is depopulated through two competing decay channels, the first one radiative with a rate $1/\tau$ and the second one, non-radiative, through a thermal quenching mechanism. E is the thermal energy barrier and α expresses the balance between the two channels. Fast scintillation can therefore be obtained for intrinsically slow transitions at the expense of a loss in light output. This is the case of Lead Tungstate (PWO) with a low light yield but 10 ns decay time at room temperature to be compared to a 25 times larger light yield but 6 μs decay time at 80°K [6]). More details about thermal quenching will be given in Sect. 3.2.

Special attention must be given to afterglow, which limits the counting rate of scintillation detectors. Afterglow is a phosphorescence mechanism induced by the thermal release of charge carriers from traps. These carriers will eventually recombine on luminescence centres, causing a delayed luminescence, which can reach several percent after 1 ms for NaI(Tl) or CsI(Tl). Other crystals have a much lower level of afterglow, such as BGO (Bismuth Germanate): 0.005% after 3 ms, and CsF (Cesium Fluoride): 0.003% after 6 ms [7].

3.1.2.3 Radiation Hardness

Inorganic scintillators have in general a good stability of their scintillation properties even in the presence of intense ionizing radiation environment. This property is crucially important for detectors in space, oil well logging and high-energy physics experiments at high luminosity accelerators. The radiation hardness of the scintillation mechanism is related to the strong electrostatic field of the crystal

lattice, which shields the luminescent centres. However, the transport of light through the crystal may be affected by the production of colour centres, which absorb part of the scintillation light on its way to the photodetector. The formation of colour centres results from the trapping of electric charges by crystal structural defects or impurities and is therefore directly correlated to the quality of the raw material and of the growth technology. A large effort is needed to purify the raw materials to the required quality and to minimize the amount of structural defects during the crystal growth. However, in some cases, a specific doping of the crystal has proven to be an efficient and economical way of significantly increasing the radiation hardness [8].

3.1.3 Scintillator Requirements for Various Applications

The choice of a scintillator depends on the energy of the ionizing radiation to be detected and on constraints specific to the application. It is therefore tailored to the user requirements considering the relative importance of several parameters, such as density, light yield, scintillation kinetics, emission spectrum, radiation hardness. Ruggedness, hygroscopic behaviour and production cost are also important parameters. In practice, it is impossible to find a scintillator, which combines all the most desirable properties. Besides a number of industrial applications for process control, container inspection, thickness gauging, ore processing and oil well logging a large fraction of the scintillator market is driven by X-ray and γ-ray spectroscopy in the following areas:

- High and medium energy physics particle detectors;
- Astrophysics and space applications;
- Spectrometry of low energy γ-quanta;
- Medical imaging;
- Safety Systems and Homeland Security.

The most important user requirements for each of these categories are detailed below.

3.1.3.1 High and Medium Energy Physics Particle Detectors

Scintillators are used in High Energy Physics for compact, high precision, homogeneous electromagnetic calorimetry. The purpose is to measure with the highest achievable precision the energy of electrons and photons, generally the decay products of unstable heavier particles, over the widest possible energy range.

The first important requirement is a high density material. High energy implies a high particle multiplicity of the particle collisions and requires a high granularity with good lateral containment of the particle initiated showers in order to minimize overlapping showers and to ease event reconstruction. A small Moliere radius is

therefore required, which will also improve the electron identification and allow π^0 rejection with good efficiency in high multiplicity events. More generally, a high stopping power is mandatory to longitudinally contain high energy showers in a reasonable volume and cost (typically 20–25 X_0 are needed in high energy calorimeters to contain at least 95% of the shower). Total lateral and longitudinal containment of the showers is a prerequisite to minimize leakage fluctuations and to achieve good energy resolution.

Fast scintillation is also an important parameter. In the search for rare events, and at hadron colliders, one operates at high collision rates, which requires a short time response of the detectors. Decay times of the order of the bunch crossing time (typically 25 ns) or even less are necessary. Only optically allowed (inter-configuration) transitions (like the transition 5d $\rightarrow$ 4f for Ce^{3+}), cross-luminescence, which is intrinsically fast and temperature independent as observed in Barium Fluoride (BaF_2), and strongly quenched intrinsic luminescence (as for PWO) can give rise to a fast light signal.

The demand for a high light yield is less stringent at high energy (GeV range) than at low energy (MeV range), because of the high number of scintillation photons produced even by a poor scintillator, allowing a good signal detection above the electronic noise. Such low light yield scintillators can therefore also be used for calorimetric applications in magnetic spectrometers due the rapid development of silicon photomultipliers (SiPM), with a gain comparable to photomultiplier tubes (PMT) and with the additional advantages of being very compact and immune to strong magnetic fields.

However, the high track density and event pile-up at high luminosity colliders pose serious challenges for physics event reconstruction and analysis. At the Large Hadron Collider (LHC) at CERN up to 40 pile-up events and more can be produced at each bunch crossing at the design luminosity of 2.10^{34} cm^{-2} s^{-1}, which will reach 200 pile-up events when the luminosity will be increased to 10^{35} cm^{-2} s^{-1} at the High Luminosity LHC [9]. For a collision region of about 10 cm (bunch length) the collisions will be distributed over 300 ps (Fig. 3.4 left panel). Precise temporal association of collision tracks or jets would help mitigate the pile-up. If this can be done for charged particles at high transverse momentum with particle tracking detectors this approach will be much more difficult in the forward-backward region and even impossible for neutral particles. In this case only time-of-flight (TOF) techniques can be applied as shown on the right panel of Fig. 3.4, where the two crossing bunches are symbolized by blue and red bars while their overlapping area is represented by a white bar. Events generated in the centre of the detector ($z = 0$) will generate tracks arriving at the same time in the forward and backward regions. On the other hand, events generated at any time off-centre of the bunch-overlapping region will exhibit a TOF difference for the tracks generated in the forward and backward regions, as shown on Fig. 3.4 (t_2–t_4, t_5–t_7, t_8–t_{10}). A mitigation factor of one order of magnitude necessitates a TOF precision of at least 30 ps [9].

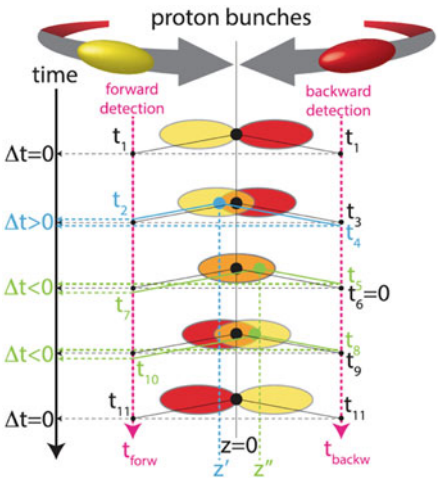


Fig. 3.4 Schematics of bunch crossing and TOF in the forward and backward directions of particles generated by events created in different positions of the overlap region

Excellent timing resolution is therefore needed. It can be shown [10] that it is related to the time density of the detected scintillation photons in the leading edge of the scintillation pulse, which is given by the following formula:

$$\sigma_t \approx \sqrt{\tau_r \tau_d \Big/ N_{pe}}$$

where τ_r and τ_d are the scintillator rise time and decay time respectively and N_{pe} is the number of photoelectrons produced in the photodetector, which is proportional to the light yield of the scintillator. A high light yield is therefore mandatory to minimize the photo-statistic fluctuation influencing the time jitter of the detector. An emission spectrum in the visible region is preferred as the quantum efficiency of the majority of photodetectors is higher and the light is generally less attenuated than in the UV region and hence more easily collected.

The energy resolution of the calorimeter is affected by all possible sources of non-uniformity. The light collection in a pointing geometry of tapered crystals introduces non-uniformity due to a focusing effect through the successive reflections of the light on the lateral faces, which depends on the refractive index of the crystal. Fluoride crystals and glasses, with low refractive index (around 1.5) have smaller non-uniformities (and therefore are easier to correct) than BGO (index 2.15) or PWO (index 2.3). The material can be intrinsically luminescent if it holds luminescent molecular complexes or ions, or is doped with a scintillating activator. Intrinsic scintillators are generally preferred, as it is easier to control the light yield uniformity in long crystals. On the other hand, a controlled distribution of the doping could help correcting for the non-uniformity caused by the light collection in a pointing geometry. Furthermore, the scintillation yield should be as independent as possible from temperature. Large temperature coefficients increase the complexity

of the detector design and of the software corrections, and temperature gradients between the front and back face of the crystals introduce non-uniformity affecting the resolution.

Finally, for large scintillator volumes cost considerations are of importance. The abundance of the raw materials, the facility to purify them against the most detrimental impurities to achieve good radiation hardness, a low temperature melting point to save on the energy cost, a high growing and mechanical processing yield are all parameters, which deserve particular attention.

3.1.3.2 Astrophysics and Space

Increasingly crystal-based calorimeters are embarked on satellites to study galactic and extra-galactic X- and γ-ray sources. This requires excellent energy resolution over a wide energy spectrum, typically from a few KeV to several TeV (see for instance Fig. 2.16 of ref. [11] for a list of different space missions with their respective energy range). One major aim of these measurements is the determination of the direction of the γ-ray source. Two classes of position sensitive devices have been developed in the last decades. These designs are using continuous scintillation crystal or pixilated detector geometries [12]. The required angular resolution is achieved with multilayer calorimeters or readout schemes to provide depth of interaction (DOI) information or using coded aperture masks.

The low orbit satellites are shielded by the earth magnetic field, relaxing therefore the requirement for radiation hardness of the scintillation material. Most of the scintillation materials can be used depending on the energy range of the detected γ-radiation. However, the payload is limiting the size of such detectors and not too dense materials are sometimes selected to reduce the weight.

In the interplanetary space the sun wind from charged particles strongly influences the detecting requirements of the scintillation materials. For these missions, high radiation hardness to ionizing radiation and low level of induced radioactivity are required. The same applies to detectors for planetary missions.

The general trend is to select high light yield, fast and not necessarily ultra-dense scintillators such as CsI or YAP. The very bright $LaBr_3$ is likely to find some applications in this domain because of its excellent low energy resolution (comparable to solid state detectors). BGO is very often used in veto counters for the rejection of Compton events.

3.1.3.3 Spectrometry of Low Energy γ-Quanta

This is probably the most important application domain for inorganic scintillators. The key requirement concerns energy resolution on the photopeak. It is therefore essential to maximize the photofraction and high Z materials are clearly preferred (see Sect. 3.1.1).

The energy resolution is driven by several factors and a detailed discussion is given in Sect. 3.1.1. However, two important parameters are playing an essential role. The first one is the light yield. One contribution to the energy resolution is the statistical fluctuation of the number of photoelectrons, n_{pe}, produced in the photodetector. Therefore a high light yield will reduce this statistical contribution like $(n_{pe})^{-1/2}$.

The second parameter concerns the deviations from the linearity of response at low energy. Most crystals exhibit a non-proportionality behaviour for energies below 100 keV. The relative light yield can show either relative increase with decreasing energy, as is the case for halide crystals, or a decrease, as for the majority of oxides and fluorides. Only few crystals have an almost linear response down to about 10 keV, such as $YAlO_3$ (YAP), $LuAlO_3$ (LuAP), $LuYAlO_3$ (LuYAP), $LaBr_3$. Given that the energy loss mechanisms—photoelectric, Compton scattering and pair production—are energy dependent, the total energy deposit in a crystal detector will be a mix of these contributions varying with energies. The non-linearity affects therefore the energy resolution, as is illustrated by the examples of Lutetium orthosilicate (LSO) and Lutetium Aluminium Perovskite (LuYAP). For the same detector volume, LuYAP achieves similar energy resolution (9%@511KeV) as LSO despite a three times lower light yield [13], as a result of a more linear response at low energy, as shown on Fig. 3.5.

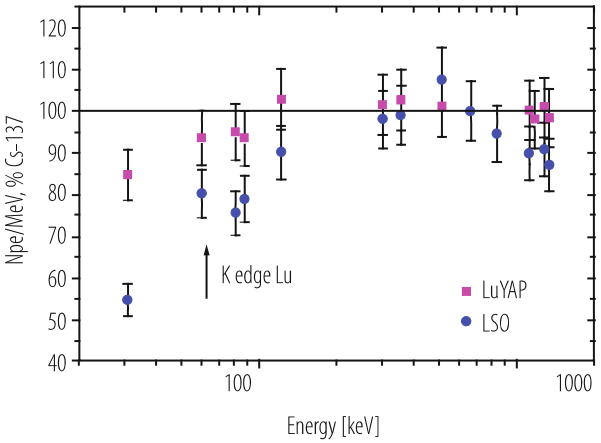


Fig. 3.5 Relative low energy response for LSO and LuYAP crystals, normalized to the ^{137}Cs energy peak (from ref [13])

3.1.3.4 Medical Imaging

Scintillators are widely used in medical imaging for X-ray radiology (digital radiography and CT scanners) and for emission tomography (PET and SPECT) with a market exceeding several hundred tons per year (see Sect. 20.1).

The choice of the scintillator for medical imaging devices is determined by the stopping power for the energy range of X and γ-rays to be considered, or more precisely the conversion efficiency. Materials with high Z and high density are favoured but the energy of the K-edge is also important as can be seen in Fig. 3.6. For low energy X-ray imaging (below 63 keV) the attenuation coefficient of Yttrium, Cesium and Iodine are quite high and crystals like YAP and CsI are good candidates for soft tissue X-ray imaging like mammography. Above the K-edge of Lu (63 keV) and Bismuth (90 keV) the situation is quite different and BGO and Lutetium based crystals are favored for bone, dental X-ray, ^{99}Tc (90 keV) SPECT and PET scanners (511 keV). Heavy scintillators have smaller thickness, reducing parallax errors in ring imagers and maintaining a good spatial resolution over the whole field of view (Sect. 7.1).

A high light yield is also mandatory for good energy resolution. Better energy resolution increases rejection of Compton events, improves the spatial resolution and the sensitivity. The sensitivity is a critical parameter as it determines the number of useful events per unit of injected dose. A higher sensitivity means a smaller injected dose or a better image contrast.

A short scintillation decay time reduces the dead time and therefore increases the maximum counting rate. In PET scanners for instance reducing the coincidence

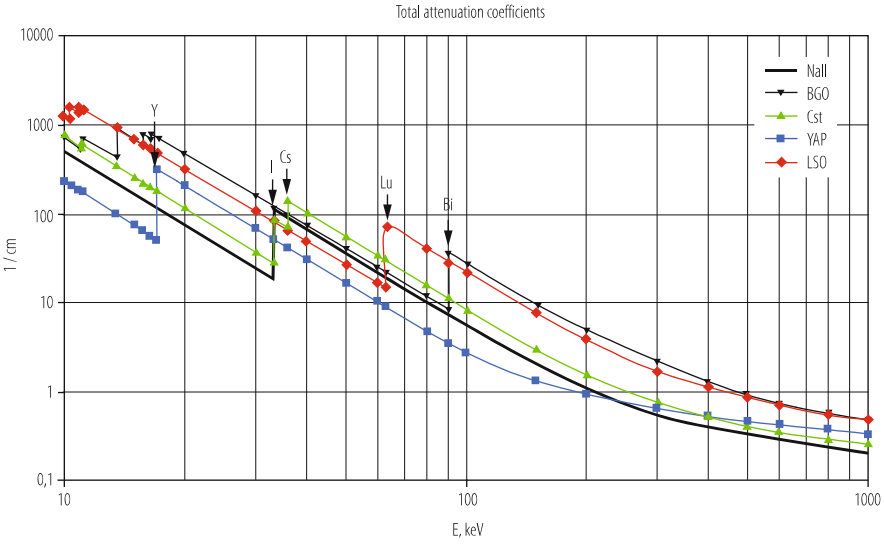


Fig. 3.6 Attenuation coefficients in several high Z materials

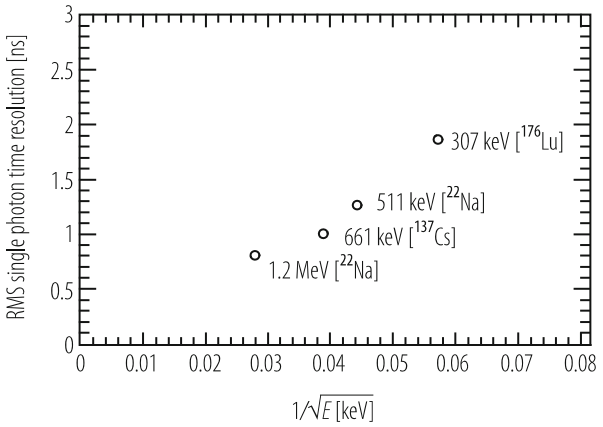


Fig. 3.7 Energy dependence of the timing resolution of a ClearPEM $2 \times 2 \times 20$ mm^3 LSO pixel coupled to an Hamamatsu avalanche photodiode (courtesy J. Varela)

window improves the signal to background ratio and increases the sensitivity and image contrast. Very fast scintillators open the way to scanners using the time-of-flight information, which helps reducing the background by selecting a narrow region of interest along the coincidence line. In the range of energies considered for medical imaging, the timing resolution is limited by the Poisson distribution of photons arrival time on the photodetector, even for bright scintillators like LSO. Figure 3.7 shows the $1/\sqrt{E}$ dependence of the timing resolution of a ClearPEM [14] detector head made of $2 \times 2 \times 20$ mm^3 LSO pixels coupled to a 32-channel Hamamatsu APD matrix, when excited by sources at different energies E.

Commercial PET scanners achieve about 500 ps FWHM coincidence time resolution (CTR) in the difference of detection time of the two 511 KeV gamma rays resulting from the positron annihilation. This allows a significant image quality improvement particularly for over-weighted patients. Ideally, one would like to achieve 100 ps FWHM CTR resolution, which would correspond to a centimetre resolution along the line of response (LOR) corresponding to the coincidence detection of the two gamma rays. It improves by an additional factor 5 the image signal-to-noise ratio. Thus a TOF-PET system with 100 ps CTR can either give a five times shorter examination time of the patient or a five times lower radiation dose at constant image quality.

As mentioned in Sect. 3.1.2.2, in first approximation (assuming single photon detection) the CTR for a scintillators with a scintillator rise time τ_r and a decay time $\tau_{d,}$ is given by:

$$CTR \propto \sqrt{\frac{\tau_r \tau_d}{N_{phe}}}$$

where N_{phe} is the number of photoelectrons readout from the crystal. Clearly, there is a premium for a high photon rate in the leading edge of the scintillation pulse, a high light yield as well as a short rise and decay times for improving the CTR.

3.1.3.5 Safety Systems and Homeland Security

Scintillators are used in three main types of equipment related to safety and homeland security: express control of luggage and passengers, search for explosive materials and remote detection of fissile materials.

Luggage inspection requires the highest possible throughput to quickly identify a suspect luggage in a few cubic meter large container moving across the inspection device. The spatial resolution is determined by the need to quickly localize and identify the suspect object in a large container. Fast scintillation kinetics with no afterglow is therefore the most important parameter.

For the remote detection of explosives the most attractive methods are based on the detection of natural or induced characteristic neutron and γ-rays under activation by a neutron source, either with fast neutrons from the ^{252}Cf radioisotope or fast-thermal neutrons from a pulsed electronic neutron generator. Neutrons initiate nuclear reactions in some elements, some of them producing characteristic γ-rays. Plastic explosives for instance are generally rich in nitrogen. The nitrogen (n,γ) reaction has a cross section of 75 mb and produces a characteristic γ-ray of 10.83 MeV.

For such applications, the most important scintillation crystal parameters are: high stopping power to improve the detector sensitivity; high light yield to improve the detector energy selectivity; fast scintillation decay time to allow time-of-flight analysis with pulsed neutron generators to increase the signal to noise ratio. Good stability of the scintillator parameters under ionizing and neutron irradiation allows the use of strong activation sources for a better sensitivity.

Remote detection and fissile materials warhead inspection has been for a long time restricted to the detection of neutrons, as the γ-channel would have easily revealed secret characteristics of the nuclear device. This has changed recently and opens new possibilities to detect the radiation emitted by Nuclear Explosive Devices (NED) based on enriched uranium or plutonium. The most useful energy range to detect fissile material is $E_\gamma \geq 3$ MeV because of (1) the absence in this range of natural radioactive sources and therefore an acceptable signal to background ratio; (2) the high penetration power of these energetic γ-quanta making the deliberate concealment of the intrinsic NED radiation more difficult.

Here, the most important parameters are sensitivity to allow detection at large distance (at least several meters) and good background rejection. High stopping power (and therefore high density) is mandatory. However, the crystals should be made from materials with very low natural radioactivity, which restricts the choice of heavy materials to the ones with no unstable isotopes. As the counting rates are usually low, there is no need for ultra-fast scintillators. A phoswich geometry based on two different crystals on top of each other can be an attractive solution for

improved low energy background rejection. A first thin scintillator layer detects (and rejects) the low energy background activity, whereas a thicker layer on the back will be mainly sensitive to the 3–10 MeV range of interest. The two scintillators must have different emission wavelength and/or decay time for a good identification of the hit source.

3.1.4 Organic Material, Glass and Condensed Gases

There is a particular class of scintillators, which does not require a regular lattice to produce scintillation light when excited by ionizing radiation. These are organic solid and liquid materials, condensed gases as well as scintillating glasses. A common feature of all these materials is that scintillation (also called fluorescence in this case) results from a direct excitation of a molecule and does not involve the transport of the excitation energy through the material. As the molecule is directly excited and the coupling with the host material is minimal, the fluorescence decay time is solely determined by the quantum numbers of the excited and ground states. If properly chosen the molecule will emit between two singlet states giving rise to a fast emission (usually not more than a few ns).

Different material combinations can be engineered, in particular in plastic scintillators, to meet specific requirements. The most popular one concerns wavelength shifters. Binary or even ternary solutions of different fluors can be dissolved in a plastic base containing aromatic molecules. After excitation by ionizing radiation, these aromatic rings will relax the stored energy by emitting UV photons. Properly chosen additional fluors can absorb these photons and reemit them at longer wavelength, e.g. to better match the quantum efficiency of a photodetector. As there are only energy transfer and no charge transfer mechanisms involved, the whole process is very fast.

Plastic scintillators can be easily machined in any shape, including in the form of fibres, one important advantage. However, these materials are intrinsically light (density around 1–1.2 g/cm^3) and therefore are not suitable for homogeneous calorimetry. They find a number of applications in sampling calorimetry and tracking. More information can be found in ref. [15].

3.2 Scintillation and Quenching Mechanisms in Inorganic Scintillators

3.2.1 The Five Steps in Scintillation Process

In contrast to luminescence (such as in lasers), where the excitation source is tuned to the energy levels of the luminescent centres, scintillation is the result of a

complex chain of processes, each of them characterized by a specific time constant and efficiency factors [16]. This is summarized in Fig. 3.8, where the valence and conduction bands of an insulator with a bandgap width E_g (forbidden band) are represented. The upper level core band (energy E_c and bandwidth ΔE_c) is also shown.

The sequence of processes is shown as a function of time and can be qualitatively divided into five main phases:

- The first one is the energy conversion phase and the subsequent production of primary excitations by interaction of ionizing particles with the material. For an incident particle energy in the keV range or higher, the excitations are essentially deep holes h created in inner core bands and hot electrons e in the conduction band. Subsequently, on a very short time scale (10^{-16}–10^{-14} s), a large number of secondary electronic excitations are produced through inelastic electron-electron scattering and Auger processes with creation of electrons in the conduction band and holes in core and valence bands. At the end of this stage, the multiplication of excitations stops. All electrons in the conduction band have an energy smaller than $2E_g$ (e-e scattering threshold) and all holes occupy the valence band if there is no core band lying above the Auger process threshold (general case).
- The second stage is the thermalization of electronic excitations through a phonon coupling mechanism with the crystal lattice, leading to low kinetic energy electrons in the bottom of the conduction band and of holes in the top of the valence band. This thermalization phase takes place in the sub-picosecond range, typically between 10^{-14} and 10^{-12} s.
- The next stage, between 10^{-12} and 10^{-10} s, is characterized by the localization of the excitations through their interaction with stable defects and impurities of the material. For example, electrons and holes can be captured by different traps or self-trapped in the crystal lattice. Excitons, self-trapped excitons, self-trapped holes (V_K centers) can be formed with emission of phonons. Localization of excitations can be sometimes accompanied by displacements of atoms (defect creation, photo-stimulated desorption).
- The transfer of excitations to the luminescent centres through the sequential capture of charge carriers or different energy transfer mechanisms takes place during the following 10^{-10} and 10^{-8} s.
- Finally, the radiative relaxation of the excited luminescent enters produces the light signal with an efficiency and time structure, which is given by the quantum selection rules of the transition. Parity allowed transitions with more than 3 eV energy gaps are generally preferred as they give rise to fast luminescence. However, smaller energy gaps (2–3 eV) are likely to favour higher light yield, as discussed in Sect. 3.1.2.2.

The scheme depicted in Fig. 3.8 describes the scintillation mechanisms in the case of ionic crystals with simple energy structures. However, important groups of scintillators exhibit a more complicated band structure.

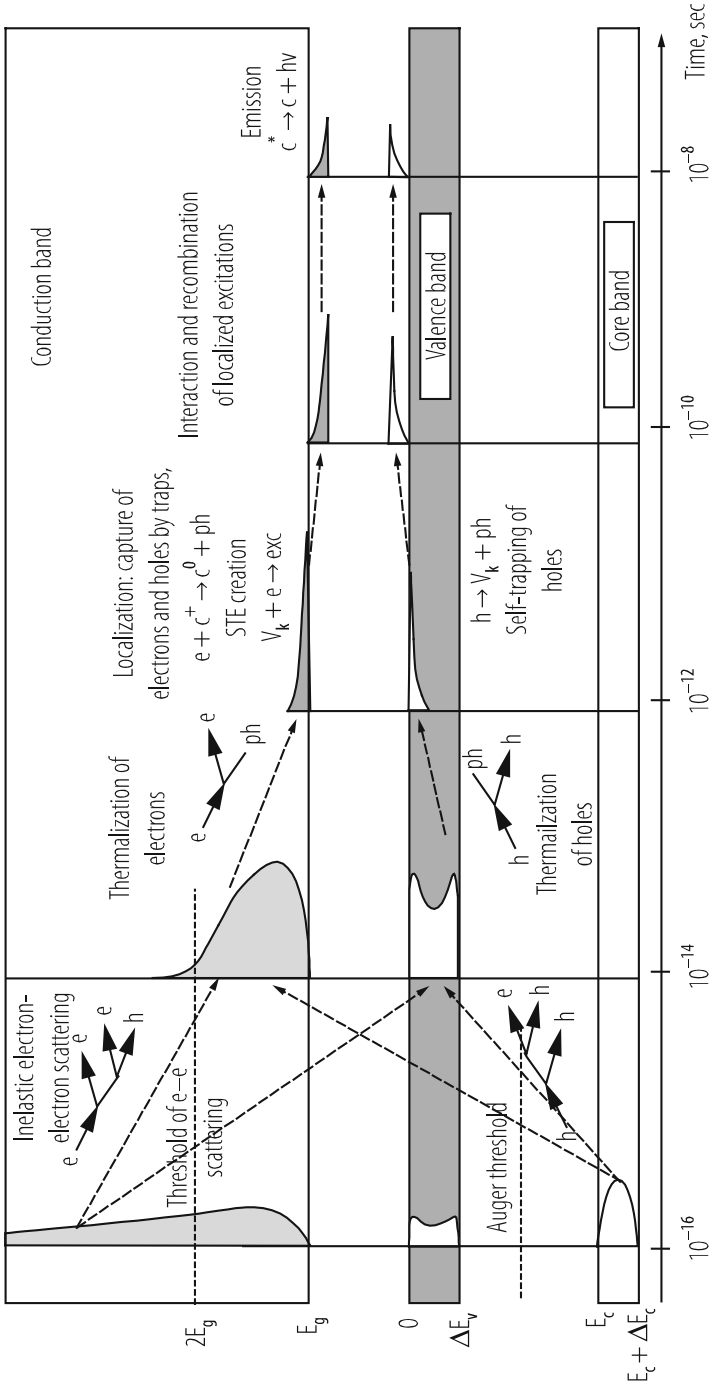


Fig. 3.8 Relaxation scheme for electronic excitations in an insulator: e, electrons; h, holes; ph, phonons; hν, photons; V_k, self-trapped holes; STE, self-trapped excitons; c^n, ionic centers with charge n. Density of states is represented by grey and white areas for electrons and holes respectively (courtesy A. Vasiliev)

One such case are so-called cross-luminescent materials, of which one well-known example is Barium Fluoride (BaF_2). Such systems are characterized by a specific configuration of the energy bands, such that the width of the forbidden gap (between the valence and conduction bands) is larger than the energy gap between the uppermost core band (5pBa in the case of BaF_2) and the bottom of the valence band. When a hole produced in this core band recombines with an electron of the valence band there is not enough energy available to eject an Auger electron from the valence to the conduction band. The core-valence transition can therefore only be radiative giving rise to a scintillation in the UV, which is usually very fast (sub-nanosecond).

3.2.2 Scintillation Efficiency

The overall scintillation efficiency η is generally given by the product of three terms:

$$\eta = \beta \cdot S \cdot Q \tag{3.10}$$

where β represents the conversion efficiency for the production of electron-hole pairs, S the excitation transport efficiency, including thermalization of electric carriers, localization and transfer to the luminescent centre, and Q is is the quantum efficiency of the radiative transition of the luminescent centre. If we consider, as discussed in Sect. 3.1.2.2, the number of 140,000 ph/MeV as an upper limit for the scintillation yield of an ideal scintillator with an emission peak around 600 nm the maximum scintillation efficiency is less than 30%. In reality, for the majority of existing scintillators it is less than 5%, mostly because of important losses during the thermalization and transport process.

At the end of the first phase of inelastic scattering the holes and electrons have reached an energy below the Auger and ionization thresholds respectively. Their thermalization to the top of the valence band for holes and to the bottom of the conduction band for electrons can only take place by heat dissipation through coupling to the phonon modes of the lattice. This is an unavoidable part of energy loss for the scintillation process. The energy gap between these two thresholds being of the order of $2.3E_g$ for ionic crystals one concludes that an ideal scintillator cannot convert more than 43% of the absorbed energy into light.

Another important loss is related to the transfer of the excitations to the luminescent centres. A frequent channel of excitation for acceptors is a charge transfer process with a sequential capture of charge carriers. In Ce^{3+}-doped crystals, the hole is first captured with its capture probability strongly depending on the position of the Ce^{3+} ground level (4f) in the forbidden band gap. In cerium-doped oxides and halides, this level is usually lying very low in the gap close to the top of the valence band, and these systems can lead to very efficient scintillation (LSO, LuAP, $LaCl_3$, etc.). On the other hand, Ce^{3+}-doped fluoride crystals cannot exhibit

very high light yield because the Ce^{3+} 4f is lying around 3–4 eV above the valence band, which strongly reduces the hole capture probability.

It is also important to avoid the delocalization of electrons from the activator excited state to the conduction band. This is achieved if the energy gap ΔE between the radiating level of the doping ion and the bottom of the conduction band is large enough. If $\Delta E >> kT$, or the radiative decay $\tau_\gamma << \tau_d$, where the delocalization time $\tau_d \approx (1/S)\exp(-\Delta E/kT)$, with S—the frequency factor, k—the Boltzman constant, and T—the temperature, the scintillation yield is not strongly dependent on the temperature. In the reverse case one can expect a reduction of the scintillation yield when the temperature increases (temperature quenching). Similarly, when the ground state is located in or very close to the valence band, the hole is weakly trapped and can be easily delocalized to the valence band.

Besides these different processes, a number of competing channels can limit the probability of charge carrier capture by the luminescent centres. Impurities or ions in the lattice can act as specific killer ions and compete with active ions for the capture of charge carriers and/or interact with them, inducing severe limitations in scintillation efficiency. For example, in cerium-doped crystals the presence of ions or molecular groups with two or more stable valence states is generally to be avoided. This is due to the fact that cerium has two stable valence states, Ce^{3+} and Ce^{4+}, but Ce^{3+} only gives rise to luminescence. If a possibility exists for Ce^{3+} to transfer one electron to these killers it will transform into Ce^{4+} and no longer scintillate. This is the case for Ce-doped tungstates and vanadates, which do not exhibit cerium scintillation because of such Ce-W and Ce-V interactions. For the same reason the good electron acceptor Yb^{3+} severely quenches the Ce^{3+} scintillation.

Self-trapping is also a very frequent source of efficiency loss in insulating materials. Indeed, some of the electrons and holes can be trapped by impurity or crystal defect related acceptors and cannot excite directly luminescent centres through sequential capture. If the trap is very shallow it will quickly release the charge carriers and will slightly delay scintillation. However, in deep traps strong quenching of the fast luminescence components is observed. Very long components in the fluorescence decay appear when the temperature is raised to the point, where trapped electrons can be released by thermal energy (glow peaks).

The interaction between closely spaced electronic excitations (in a few nanometre range) may lead to luminescence quenching, also-called local density-induced quenching. For electronic excitations created through the different mechanisms of photon absorption, the probability to produce excitations at such short distances is very low if the excitation source has a limited intensity. On the contrary, secondary electronic excitations created by inelastic scattering of photoelectrons or Auger decay of core holes can be quite closely spaced. In these clusters of high local e and h density, the interaction between excitations can modify their localization and can even create defects in crystals. In addition, these clusters can excite closely spaced luminescent centres, which can interact with each-others, giving rise to faster and

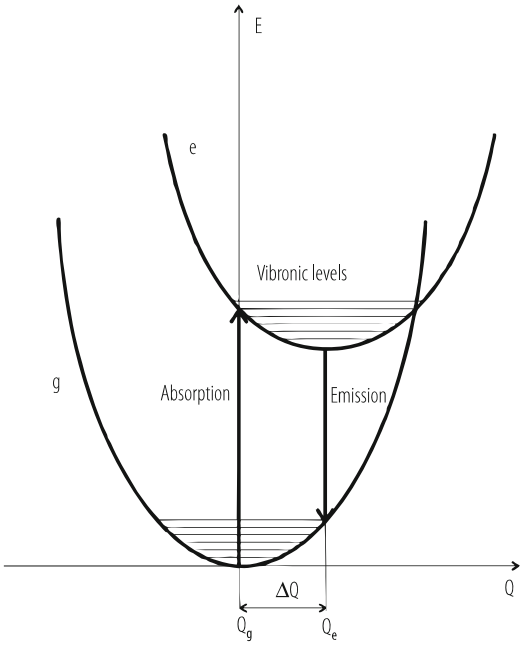


Fig. 3.9 The configurational coordinate diagram. The energy E is plotted versus the coordinate Q (configurational coordinate in the lattice). The ground state g and one excited state e are represented by potential curves with offset ΔQ. Absorption and emission transitions are indicated

non-exponential decay time and total or partial luminescence quenching. The first evidence of such effect was observed in CeF_3 [17].

Another type of thermal quenching can occur related to electron-phonon coupling. The different electronic configuration of the ground and excited states of the activator generally induces an exchange of phonons and the relaxation of the position of the activator ion when it is excited. As a result, the emission transition from the relaxed excited state is shifted towards lower energy than the absorption transition. This is the well-known Stokes shift illustrated in Fig. 3.9. The Stokes shift is a measure of the interaction between the emitting centre and the vibrating lattice. The stronger the electron-phonon coupling the larger the Stokes shift. For weak coupling, the potential curves are not significantly shifted and the emission spectra show narrow lines (case of f-f transitions of rare earth ions). In the case of intermediate coupling for which the parabolas are weakly shifted, vibronic spectra of broad emission lines are observed reflecting the progression in stretching vibration of the luminescent ion (case of uranyl pseudo-molecules in oxides, like UO_2^{2+}).

In the case of strong coupling (shown in Fig. 3.8) the relaxed excited state may decay non-radiatively to the ground state if the temperature is high enough to allow the excitation to reach the crossing of the two parabolas.

In practice, the relevant parameter is the light yield efficiency Y, which is the product of the scintillation yield η by the light transport and collection efficiency η_{col} to the photodetector. A number of parameters influence η_{col}: the crystal shape, its optical transparency to the scintillation wavelength, the presence of scatters and

different optical defects in the bulk of the crystal, the surface state and wrapping conditions of the faces of the crystal, the coupling face to the photodetector, the surface matching between the coupling face and the photodetector, the crystal index of refraction. Heavy scintillators generally have a high index of refraction (larger than 2 in many cases) and the light collection efficiency is limited to 10–30% for the majority of existing detectors. New approaches based on nanostructured surfaces, in particular photonic crystals, are presently being explored [18]. Significant light extraction gains of more than 50% have been obtained as well as a strong reduction of the photon transit time spread in the crystal associated to the higher extraction probability of the photons at their first hit on the coupling face to the photodetector (reduction of multiple bouncing) [19].

The fact that some self-activated scintillators, like $PbWO_4$, exhibit fast room temperature scintillation in the ns-range is only the consequence of a luminescence quenching mechanism competing with the radiative relaxation of the excitation. In this case the decay is non-exponential, which is a common signature of temperature quenched scintillators.

3.2.3 Response Linearity and Energy Resolution

The ultimate energy resolution (FWHM) of a perfect scintillator based detector is given by the well-known Poisson law:

$$R_{\text{lim}} = 2.35\sqrt{\frac{1 + v(PD)}{N_{\text{pe}}}} \tag{3.11}$$

where $v(PD)$ is the variance of the photodetector gain and N_{pe} is the number of photoelectrons emitted by the photodetector. As the number of photoelectrons is proportional to the number of photons N_{ph} produced by the scintillator, the resolution should be driven by the photostatistics of the scintillator light production. However, several other factors contribute to the practical resolution R:

$$R^2 = R_{\text{lim}}^2 + R_{\text{inh}}^2 + R_{\text{tr}}^2 + R_{\text{np}}^2 \tag{3.12}$$

where R_{inh} reflects homogeneities of the crystal, inducing local variations of the scintillations efficiency, R_{tr} is related to the light transport and collection by the photodetector and R_{np} is a factor of non-proportionality, which accounts for the fact that for some scintillators, the number of emitted photons is not strictly proportional to the incident energy.

Non-linear response has been first reported for NaI(Tl) and CsI(Tl); the response per unit deposited energy decreases continuously from X- and γ-rays to electrons, protons, α particles, and fission fragments. Moreover, this trend is strongly correlated with the ionization density dE/dx [20]. In other words, the response of

a scintillator depends not only on the total amount of energy but also on the mechanisms of the energy deposit. There is common agreement that this is related to the saturation of response of the luminescent centres in the presence of a high density of charge carriers. This is parameterized by Birks law, which postulates a non-radiative relaxation of excitons interacting with each others in the case of high ionization density:

$$N_{\mathrm{ph}}\left(\frac{\mathrm{d}E}{\mathrm{d}x}\right) = \frac{N^0_{\mathrm{ph}}}{1 + a_{\mathrm{B}}\frac{\mathrm{d}E}{\mathrm{d}x}} \tag{3.13}$$

where $N^{\circ}{}_{\mathrm{ph}}$ is the light yield in the absence of saturation, N_{ph} is the actual light yield and a_{B} is the Birks parameter.

When combining the $1/\beta^2$ ionization density increase for low energy particles of decreasing velocity β (Bethe Bloch formula) with the Birks saturation law one obtains the typical scintillator non-linear response at low energy as illustrated on Fig. 3.10 in the case of NaI(Tl) [21].

It remains, however, to be explained why some scintillators are more affected by this saturation effect than others.

Each of the steps of the conversion process described in the previous section can be characterized by a certain degree of non-linearity. It seems, however, that last stages of thermalization and capture are the most affected by non-linear phenomena. Indeed, as long as the kinetic energy of electrons and holes is large relative to the bandgap E_{g} the excess energy will be used to produce secondary e-h pairs and this energy conversion process is intrinsically linear. On the other hand the stability of the thermalized excitons in its crystallographic environment is very much dependant on the energy band structure of the material as well as on the density of luminescent centres or defects. This stability is related to the correlation distance between the electron and hole, which is energy and temperature dependant.

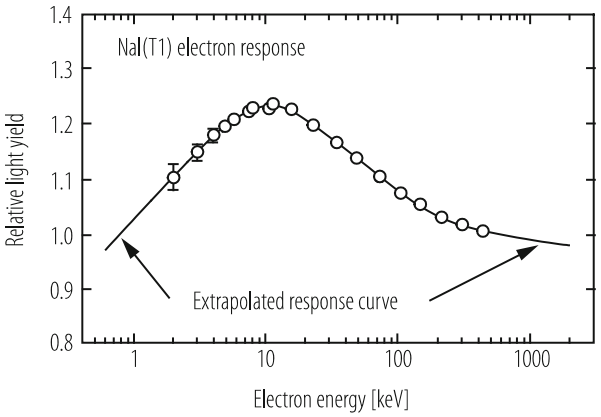


Fig. 3.10 Measured electron response for NaI(Tl) scintillator (from ref [21])

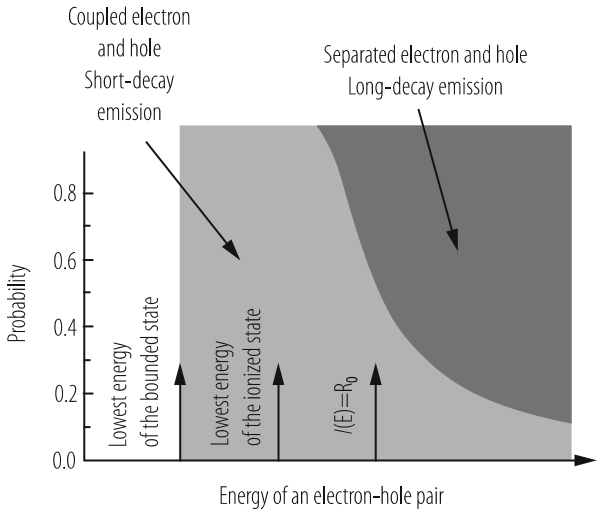


Fig. 3.11 Probability of binding or separation of an e-h pair as a function of energy (courtesy A. Vasiliev)

Two competing recombination processes can take place, both being intrinsically non-linear with energy as shown on Fig. 3.11, and inducing therefore a non-linear energy response of the scintillator. The first one is the self-trapping of the exciton in the vicinity of a luminescent centre which decreases rapidly with the e-h pair energy. The second one is the direct capture of the separate electron and/or hole by defects or luminescent centres and increases with the kinetic energy of the electron and hole.

The energy threshold between these two mechanisms is related to the correlation distance R_0 between the electron and hole, which is temperature dependant. As a result, the energy dependence of the scintillator response to thermalized e-h pars is strongly non-linear as shown on Fig. 3.10, which also shows the influence of the defects (crystal quality) on the excitation transfer efficiency to the luminescent centres.

The quantitative link between the low energy non-linearity of the scintillator response and the deviation of its energy resolution from the predicted counting statistics is far from being fully understood. It has however its seed in the fact that for the same total amount of deposited energy both photons and electrons release this energy in a number of quanta over a large energy range and that the light response for each of these quanta has different proportionality constants as a function of energy. The event-to-event variation of this cascade process induces a spread in the energy response, which deteriorates the energy resolution.

This is obvious in the case of Compton scattering. In a detector of a finite size, the events in the photopeak result from the sum of true photoelectric events and of events having undergone single or multiple Compton scattering interactions all

contained in the detector block. The total energy deposited in the detector block is the same whether it results from a single or multiple interactions. The light response may however differ due to the non-linear response of the scintillator. As a result, the event-to-event statistical variation of the energy deposition mechanism induces a broadening of the resolution.

As pointed out in ref. [22] one would expect an improvement of the resolution by reducing the detector size, as the fraction of fully contained Compton events decreases and consequently the proportion of true photoelectric events increases. This is actually not the case, because photoelectric events may result at the atomic scale from a complex cascade mechanism. Indeed, the photoelectric interaction of an X- or γ-ray produces a mono-energetic electron from one of the inner shells of the atoms of the absorber. However, this electron can be ejected not only from the K shell but also from a L or even a M shell (although the cross-section rapidly decreases for higher core levels) of the different atoms of the crystal. In the sequence of photon detection, recoil electrons with different energies are produced, each carrying the incident photon energy minus the binding energy of the shell, from which it has been ejected. Moreover, the deep hole produced in the inner shell will be filled by an electron from outer shells, which in turn will be replaced by electrons rom even lower bound shells through a cascade of relaxation events, each of them producing an X-Ray or an Auger electron converting in the crystal following the same mechanisms. Figure 3.12 depicts a part of this cascade process for LSO crystals, commonly used in medical imaging cameras.

Finally, the recoil electrons, as well as all charged particles detected in a scintillator, slow down through a sequence of energy transfers to the absorber with a progressively increasing ionization density.

The energy resolution of calorimeters used in high or medium energy physics is generally parametrized as a function of energy according to the following formula:

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c \tag{3.14}$$

where a is the statistical term, b the noise term and c a constant term, which takes into account all the systematics (intercalibration error, temperature effects, light yield non-uniformity in the crystal, shower leakage, etc.).

At high energy, the constant term is predominant and it requires a challenging engineering effort to reach the sub-percent level for large detector systems with tens of thousands of channels. This has been achieved for the LEP L3 BGO calorimeter (with 12,000 crystals) with a high energy resolution of 1% and in the LHC CMS PWO calorimeter (77,000 crystals) with a constant term of better than 0.5%.

At lower energy, the electronic noise plays an increasing role. The noise contribution, which is energy independent, contributes therefore to the *relative* energy resolution (3.14) as $1/E$.

An interesting example is given in Fig. 3.13 for two heavy Lutetium based crystals popular for medical imaging devices, LSO and LuYAP. In spite of a light

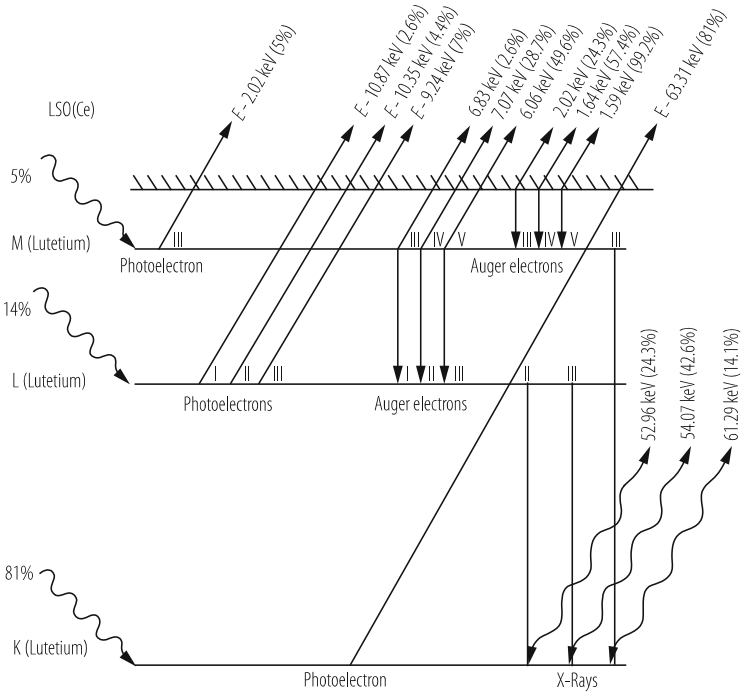


Fig. 3.12 Electron cascade following photoelectric absorption in LSO Crystal. *E* refers to the photoelectrically absorbed photon energy (ref. [22])

yield nearly three times lower LuYAP achieves a comparable energy resolution than LSO because of a much more linear behavior at low energy (see Fig. 3.4).

3.2.4 Scintillation Kinetics and Ultrafast Emission Mechanisms

Achieving ultimate time resolution on scintillator-based detectors requires a parallel effort on the light production mechanisms, light transport optimization to reduce the travel time spread of the photons on their way to the photodetector, on the photoconversion system as well as on the readout electronics.

As shown in Sect. 3.2.1 the radiative transition on the activator ion or on the intrinsic luminescent center only takes place after a complex relaxation mechanism of the primary electron-hole pairs that can last several nanoseconds. In this process large statistical fluctuations are therefore induced for the generation of the first scintillation photons, which influence the observed rise time. This presents an intrinsic limit to the achievable time resolution in a scintillator. It is related to the

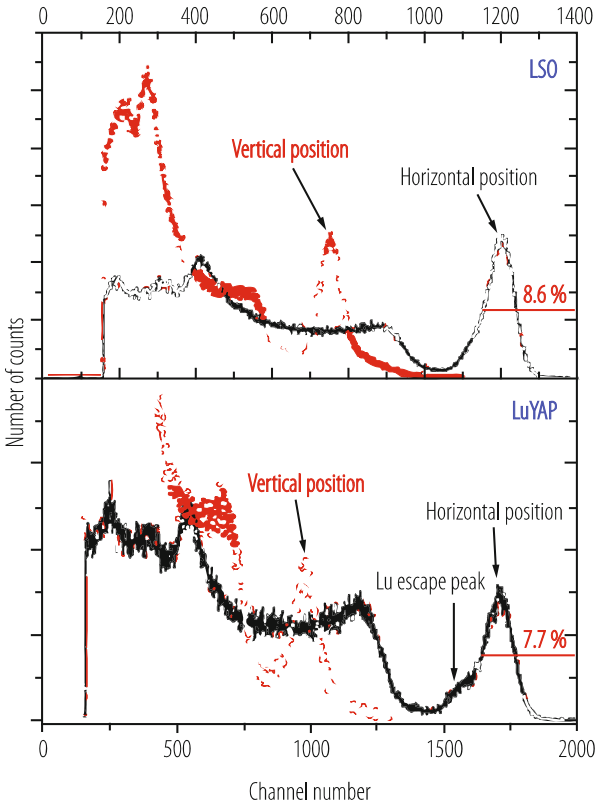


Fig. 3.13 Energy resolution for ^{137}Cs photons obtained with 2 × 2 × 10 mm^3 Ce doped LSO and LuYAP crystals measured in horizontal and vertical position. The electronic gain for LuYAP is three times higher to compensate for the lower light yield (6000 pe/MeV and 2000 pe/MeV for LSO and LuYAP, respectively, in horizontal position)

time fluctuations in the relaxation process that can be estimated to be of the order of 100 ps.

For sub-100 ps time resolution mechanisms involving the production of prompt photons need to be considered. Cherenkov emission and cross-luminescent materials can offer a solution. However, the number of Cherenkov photons from the recoil electrons resulting from a 511 KeV γ conversion is very small, of the order of 20 photons in crystals like LSO, LuAP and GSO. Moreover, these photons are preferentially emitted in the UV part of the spectrum, where the optical transmittance and the photodetector quantum efficiency are generally low. The same applies for cross-luminescent materials characterized by a reasonably fast emission (600 ps for BaF_2) which emit in the 100–250 nm spectral range. However, some transient phenomena in the relaxation process that can be possibly exploited for the generation of prompt photons. From this point of view, an interesting phase of the relaxation mechanism is the thermalization step when the hot electrons and

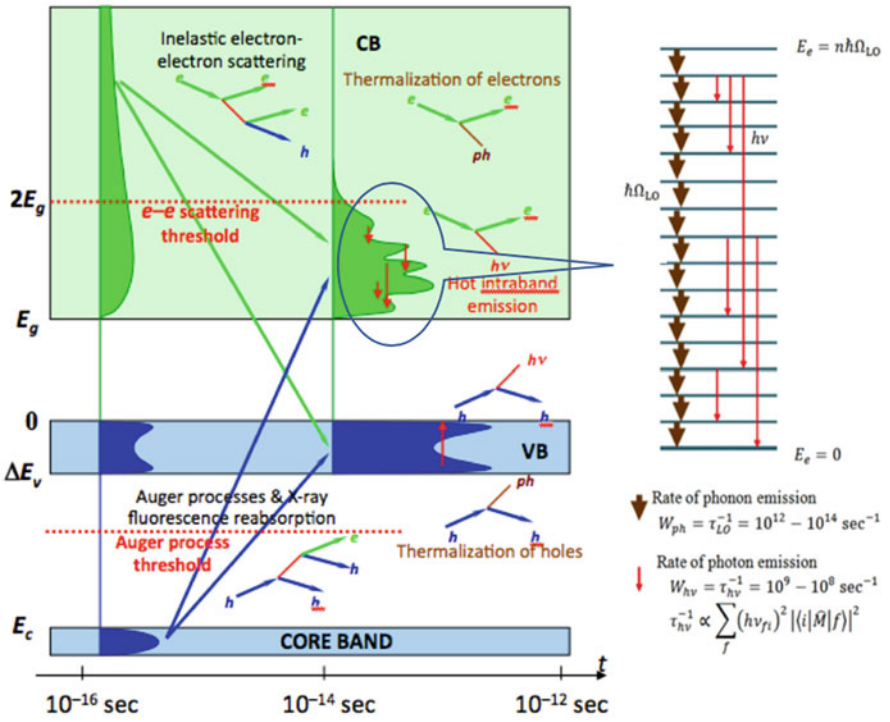


Fig. 3.14 Schematic description of the hot intraband luminescence, showing the competition of radiative and non-radiative (phonon-assisted) decay channels in the case of a non-uniform density of states in the conduction band. From Ref [23]

holes have passed the ionization threshold. The coupling to acoustic and optical phonons in the lattice is the source of hot intraband luminescence (HIBL) that could be exploited to obtain a time tag for the interaction of ionizing radiation with a precision in the picosecond range [23, 24]. This emission is rather weak but extremely fast (sub-ps) and is characterized by a flat spectrum in the visible for the electron-induced HIBL in the conduction band with an onset in the near infrared attributed to the hole HIBL in the valence band. Work is ongoing to engineer scintillators with a non-uniform density of states in the conduction and/or the valence band which may result in a more intense HIBL emission (Fig. 3.14). Already a few hundred prompt photons would suffice to significantly improve the time resolution of scintillators like LSO in the low energy (MeV) regime.

Hetero-structures based on a combination of standard scintillators (such as LSO or LYSO) and nanocrystals may be another way to produce prompt photons. Nanocrystals have gained considerable attention over the last two decades because of their excellent fluorescence properties. In such systems quantum confinement offers very attractive properties, among which a very high quantum efficiency and ultrafast decay time. Moreover, they have a broadband absorption and narrow emis-

sion, enhanced stability compared to organic dyes, and the fluorescence is tunable from the UV to the near-infrared spectral range (300–3000 nm) by nanocrystal size and material composition.

A novel route towards the realization of ultrafast timing resolution is possible with the use of colloidal CdSe nanosheets (CQwells) [24], a new class of two-dimensional materials. CQwells are solution-processed analogs to epitaxial quantum wells (Qwells). However, being synthesized in solution, they can be deposited on any substrate with arbitrary geometrical configuration. Further, a large dielectric mismatch between the inorganic CdSe CQwells and the surrounding organic environment results in much stronger quantum confinement than in epitaxial Qwells. This mismatch combined with very little dielectric screening due to the 1.5 nm CQwell thickness results in strongly enhanced exciton and biexciton binding energies of 132 and 30 meV, respectively, making both populations stable at room temperature.

The strong electron and hole confinement in one dimension and free motion in the plane has several important consequences, including strict momentum conservations rules (in contrast to quantum dots) and a giant oscillator strength transition. Momentum conservation in CQwells limits the available states for Auger transitions, reducing the recombination rate of this nonradiative channel. In addition to the enhanced exciton and biexciton binding energies, a giant oscillator transition results in radiative lifetimes that are significantly shorter than in bulk CdSe (~400 and ~100 ps, respectively). All of these properties contribute to the ultralow threshold stimulated emission (or superluminescence) with sub-ps decay time that has been observed with these CQwells (Fig. 3.15). Such systems could find

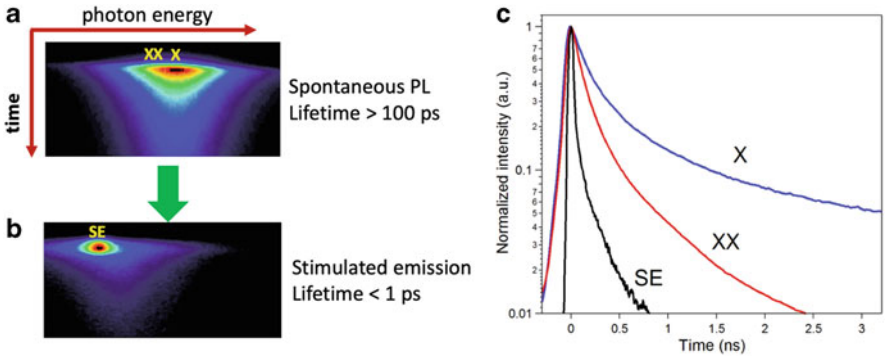


Fig. 3.15 Time-resolved spectral decay under femtosecond excitation (**a**) Streak image showing the spectral decay of exciton (X) and biexciton (XX) emission from CdSe CQwells. (**b**) Stimulated emission at an ultralow excitation fluence of $F_0 = 6\ \mu J/cm^2$, with characteristic spectral narrowing and lifetime shortening. From Ref [24]

interesting applications in ultrafast X-Ray imaging as well as providing a fast time tag in γ imaging if used in hetero-structures in combination with dense scintillators like LSO with a structuration dimension of the order of the recoil electron range, as suggested in Ref [25].

3.3 Role of Defects on Scintillation Properties and on Radiation Damage in Inorganic Scintillators

3.3.1 Structural Defects in a Crystal

The properties of a scintillator strongly depend on the structural quality of the crystal lattice. The presence of defects influences all stages of the scintillation process. They play also an important role in the light transport to the photodetector, as well as in the generation of optically active defects under radiation exposure. They continuously exchange charge carriers and phonons with the crystal lattice and are therefore in thermodynamic equilibrium with the medium. This can have a number of consequences such as reduced or enhanced scintillation efficiency if the charge carriers are channelled through these defects to non-radiative or radiative traps respectively, modification of the scintillation kinetics, afterglow, creation of perturbed emission centres, self-absorption, emission wavelength shift, radiation damage, radiation damage recovery. Depending on their size and physical nature, one can distinguish two main classes of structural defects, namely point size defects and impurities. Larger scale defects such as dislocations, twins, voids and other macroscopic defects also exist. They will not be described here, as their influence on the crystal properties is usually limited to the mechanical ruggedness and to a small extent to the optical homogeneity.

3.3.1.1 Point Size Defects

A perfect crystal is a virtual object that can only exist at absolute zero temperature. At higher temperature, a thermodynamic equilibrium is obtained by exchange of energy quanta (in the form of phonons) between the environment and the crystal lattice. Moreover, the finite dimensions of the crystal imposes conditions on the surface to compensate the electrostatic field unbalance for the atoms at the interface. This requires some level of plasticity of the lattice, which is generally achieved by a certain concentration of cation and anion vacancies. Thermodynamics imposes a relatively low concentration of such defects at room temperature, typically of the order of 10^{12} cm^{-3}. For comparison, the atomic density of the majority of known heavy scintillators is about 10^{23} cm^{-3}. In practise, the concentration of vacancies is determined by the crystal growth technology. The melt is a mixture of several chemical components, each of them with a different melting temperature

and vapour pressure, which leads to segregational evaporation of some components. Furthermore, close-to-surface vacancies can be partially compensated by absorption of ions or radicals from the surrounding atmosphere. Typically the concentration of such defects is at the level of 10^{18} cm^{-3} (10 ppm atomic) or even more. At such concentration, some collective effects can take place, leading to more complex molecular or cluster defects. Another typical point defect results from the displacement of an ion of the lattice to an interstitial position. The electrically neutral system behaves as a dipole and is called a Frenkel defect. In the case of Lead Tungstate an oxygen-based Frenkel defect is responsible for an absorption band at 360 nm and for an increased susceptibility to radiation damage.

3.3.1.2 Impurities

Impurities are ions of different nature than the constituents of the crystal lattice. They are generally introduced from imperfectly purified raw materials or by contamination, for instance from the crucible material, during the crystal growth process. Doping ions acting as luminescence activators, such as Ce^{3+} in LSO, LuAP and many other fast scintillators, can be considered as impurities with a positive role. Ions from the lattice, but in a different valence state than required by the electric charge balance, are another type of impurity. As an example, Ce^{4+} has been considered by some authors as a possible scintillation quencher in CeF_3 crystals. Two important parameters influence the way impurities can be introduced in a crystal: their electric charge and their ionic radius. If the ionic radius is close to the one of ions from the lattice, impurities can easily replace these ions, producing only a small distortion of the lattice. Isovalent ions will then easily produce a solid solution as is the case for LYSO or LuYAP when Y^{3+} ions substitute Lu^{3+} in LSO and LuAP crystals, producing locally a mixed compound of LSO-YSO and LuAP-YAP, respectively. If heterovalent impurities are introduced in the crystal their charge excess or deficit must be compensated by other impurities or by lattice ion vacancies. This mechanism can be used to suppress the detrimental role of some defects, which cannot be eliminated. A good example are the lead vacancies in PWO, which are efficient hole traps responsible for radiation damage and which can be compensated by substituting trivalent ions such as Y^{3+} or La^{3+} to neighbouring Pb^{2+} ions in the lattice.

Impurities with too large an ionic radius have generally little chance to be introduced in the lattice, whereas small ions can find interstitial positions and create strong local distortion of the crystal electronic configuration.

In practice it is difficult, or at least very expensive, to purify raw materials to the sub-ppm level. Most of the scintillators grown in good conditions have therefore an impurity concentration of about 10^{-17}–10^{-19} cm^{-3}, comparable to the concentration of point defects.

3.3.2 *Impact of Defects on Optical Properties*

Defects in a crystal influence its optical properties in a number of ways, affecting the charge carriers or the photon transport.

3.3.2.1 Charge Carrier Traps

Most point defects or impurities are electron or hole traps. They reduce therefore the transfer efficiency of charge carriers to the luminescent centres and therefore also the scintillation efficiency. For good quality crystals the density of defects (at 1–100 ppm level) is several orders of magnitude smaller than the density of luminescent centres, which is very high for intrinsic scintillators (about 10^{22} cm^{-3}) but also quite high for extrinsic scintillators, for which the activator concentration is typically at the atomic percent level. Under normal excitation conditions, it would look therefore rather unlikely that charge carriers are trapped by defects before they convert on luminescent centres. This does not take into account the charge carrier capture cross-section, which can vary by large factors for different kinds of traps. A typical example is given by the molybdenum molecular complex MoO_4^{2-}, which is a very efficient and stable electron trap with a radiative decay at 508 nm in PWO. At the level of only a few ppm it gives rise to a slow (500 ns) additional green component to the regular fast PWO emission band at 420 nm. As molybdenum is isomorphic to tungsten it can easily enter into the PWO lattice and locally produce a solid solution ($PbWO_4$-$PbMoO_4$). This slow green component is negligible if the molybdenum contamination of the tungsten oxide raw material is less than 1 ppm [26].

In some cases, the traps are non-radiative but have energy levels close enough to the valence or conduction bands so that the carriers can be released by thermal activation, eventually converting on the luminescent centres. If the trap is close to the radiative centre this thermally assisted transfer can take place directly between them without involving the valence or conduction bands. As a result, the regular emission will take place but with some delay associated with the transit of the carrier via the trap. This is the origin of the well-known afterglow or phosphorescence. When afterglow effects are undesirable, for instance for high X-ray counting rates in CT scanners, additional impurities can help opening some non-radiative relaxation channels for these traps. As an example, afterglow in $(Y,Gb)_2O_3$:Eu scintillators can be significantly reduced by the addition of heterovalent Pr^{3+} or Tb^{3+} ions to the lattice [27]. The Pr^{3+} and Tb^{3+} additives readily trap holes to form Pr^{4+} and Tb^{4+}, which compete with the intrinsic traps responsible for afterglow. This energy trapped in the Pr or Tb sites decays non-radiatively in the presence of the Eu^{3+} ion. As a consequence, afterglow emission is suppressed by one order of magnitude or more.

3.3.2.2 Defect Associated Absorption Bands

Defects have generally energy levels in the forbidden band, which reduce the optical transparency of the crystal. Small perturbations of the crystal lattice are energetically the most probable ones and give rise to a number of energy levels near the conduction and the valence bands. There is nearly a continuum of such levels, which reduces the optical transparency window of the crystal. For this reason, the shape of the optical transmission of a crystal near the band-edge is usually a good probe of its structural quality. Crystals with UV emission bands near the fundamental absorption edge are strongly affected by the optical transitions between these levels resulting in increased absorption.

Cross-luminescent crystals such as Barium Fluoride (BaF_2) are illustrative examples to demonstrate the role of impurities on the crystal properties. Their deep UV fast emission band (220 nm for BaF_2) requires a very good UV transmission to detect efficiently the light at the photodetector. Unfortunately, alkali earth fluorides are easily contaminated by oxygen and hydroxyl ions, causing strong absorption bands in the UV. A theoretical study of the charge state stability and electronic structure of O^0, O^- and O^{2-} centres in BaF_2 identified a large number of transitions from 2p to 3s and 5s states. In ref. [28] Hartree-Fock-Slatter local density discrete variation cluster calculations were made to obtain the energy levels of H_s^-, O_s^- and O_s^{2-} ions in BaF_2 crystals. Table 3.1 summarizes the optical absorption bands in the VUV and UV ranges.

As far as O^- and O^{2-} ions are concerned, the absorption bands are mainly the result of cross transitions between oxygen ions and Ba^{2+} or F^- ions, which significantly contribute to absorption around 200–240 nm.

These theoretical calculations are in good agreement with experimental results, confirming the existence of strong absorption bands overlapping the fast emission band in hydrolysed BaF_2 crystals, see Fig. 3.16.

Table 3.1 Calculated optical absorption band of H_s, O_s^- and O_s^{2-} -contaminated BaF_2 [28]

Impurities	λabs. [nm]	$h\nu$ [eV]	Cross transitions
H_S^-	209	5.9	H^- (1s) → H^- (2s)
O_S^-	230	5.4	F^- (2p) → O^- (2p,3p)
	175	7.2	F^- (2p) → O^- (3p)
	170 ≈ 175	7.0 ≈ 7.2	O^- (2p) → Ba^{2+} (5d)
O_S^{2-}	292	4.2	F^- (2p) → O^{2-} (3p)
	200	6.2	O^{2-} (2p) → Ba^{2+} (6s)
	130	9.5	O^{2-} (2p) → Ba^{2+} (5d)

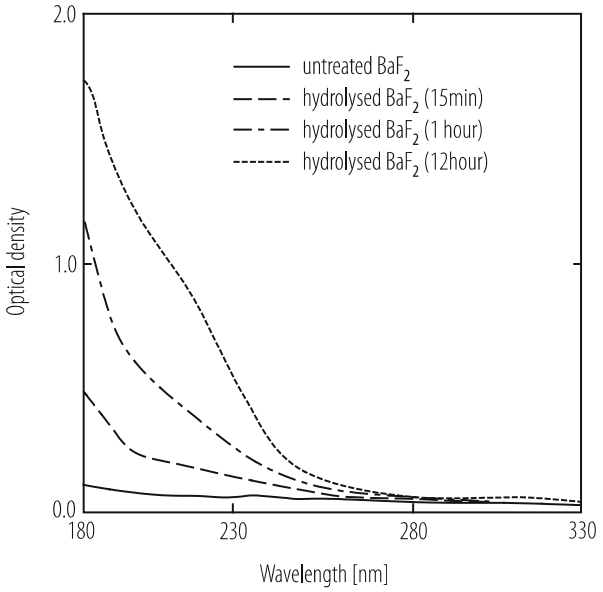


Fig. 3.16 Absorption spectra for different hydrolysed BaF_2 (ref. [28])

3.3.3 Radiation Damage

The exposure of crystals to ionizing or neutron radiation can induce a number of modifications of the crystal lattice with potential consequences for the scintillation efficiency and the light transport. These modifications can be related to pre-existing crystal defects, when exposed to a high density of charge carriers that are easily trapped producing colour centres with radiation-induced absorption bands. They can also be associated to the production of new defects by elastic or knock-on collisions of incident particles with the lattice ions resulting in a local modification of the lattice structure. Finally, heavy energetic charged particles or neutrons may produce dramatic events, such as heavily ionizing fission fragments. This last phenomenon is usually of little concern in the majority of applications, even for the new generation of high luminosity particle physics colliders, as it requires an enormous integral fluence (10^{17}–10^{18} cm^{-2}) to become significant. Indeed, it requires the formation of about 10^{17} cm^{-3} such defects to reach a 1 ppm contamination in the majority of scintillator materials. However, such defects are by nature irrecoverable and their progressive accumulation may affect parts of detectors highly exposed for very long periods of time.

The situation is different for the majority of other cases (charge trapping or ion displacement), for which relaxation processes play a fundamental role in the kinetics of damage build-up. These defects introduce a local perturbation in the crystal and do not change the main structure parameters and particularly the spatial

symmetry group. However, they locally modify the electronic configuration and affect the macroscopic crystal parameters, such as optical transmission, conductivity, thermo-luminescence properties, because these volume properties are sensitive to the microscopic structure modifications. In ionic crystals, containing anions and cations, five possible simple point defects of the crystalline structure have been observed: anion vacancy V_a, cation vacancy V_c, cation replacement by impurity ions, extrinsic atoms in inter-site positions and Frenkel type defects (anions and cations displaced to interstitial sites).

All these defects are efficient charge carrier traps and can be stabilized by capturing excess electrons or holes released by irradiation in the conduction or valence band respectively. In oxide compounds for instance, the oxygen vacancies are charge compensated by the capture of one or two electrons, which are in excess in the conduction band after irradiation. The resulting F^+: ($V_a + e^-$) and F: (V_a+2 e^-) electron centres play an important role in radiation damage effects. The captured electron or hole in these so-called recharged defects has generally a number of discrete energy levels available in the electrostatic environment of the defect and optical transitions to upper energy levels induce absorption bands in the crystal transparency window. These bands are the source of the crystal colouring under irradiation and justify the name of colour canters for these defects.

The main consequence of irradiating a crystal is to produce radiation induced absorption bands, which absorb a fraction of the scintillation light on its pathway to the photodetector. The light collected on the photodetector becomes therefore:

$$I_{\text{rad}} = \int_{\lambda} I_0(\lambda)\, e^{-(\mu_0(\lambda)+\mu_{\text{rad}}(\lambda))L} d\lambda \tag{3.15}$$

where I_{rad} is the intensity of the transmitted light after irradiation, $I_0(\lambda)$ is the intensity of transmitted light at the wavelength λ before irradiation, $\mu_0(\lambda)$ and $\mu_{\text{rad}}(\lambda)$ are, respectively, the intrinsic and radiation induced absorption coefficient at the wavelength λ and L is the mean path-length of optical photons from the emission point to the crystal exit surface. Dense and small radiation length crystals have an obvious advantage as for the same stopping power the path-length L is reduced as compared to lighter materials. Moreover, non-uniformities introduced by different path-lengths as a function of the position of the scintillation emission point are also reduced. Figure 3.17 shows the radiation induced absorption coefficient spectrum for PWO crystals as a function of the accumulated ^{60}Co dose.

At radiation levels currently experienced in particle physics detectors and in X-ray imaging devices the radiation damage only affects the optical transparency of the majority of known scintillators, but not the scintillation mechanism. One exception is CsI(Tl), characterized by an overlap of the radiation induced hole centres absorption maximum in CsI with the excitation spectrum of the Tl^+ ions. The presence of stable hole centres causes a fraction of excitations to be trapped rather than transferred to TI^+ thereby causing non-radiative losses. As a result, the efficiency of energy transfer to luminescence centres drops, decreasing the scintillation efficiency. Similarly, radiation-induced charge transfer processes can

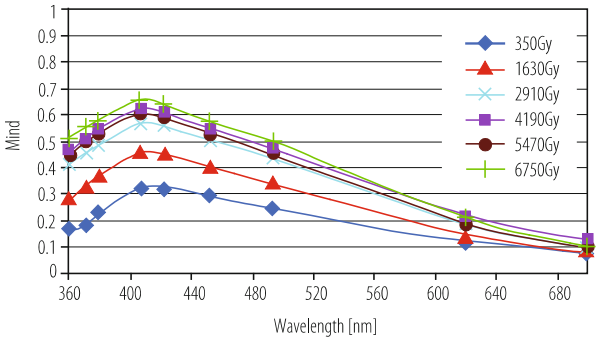


Fig. 3.17 Wavelength dependent absorption coefficient of PWO crystals as a function of the absorbed ^{60}Co dose (courtesy CMS collaboration)

modify the charge state of activator ions. This is seen for instance in some Ce^{3+} doped scintillators, such as YAP and LuYAP, when grown in vacuum or inert atmosphere, where up to several percent of the scintillating Ce^{3+} ions can be reduced under irradiation to the Ce^{2+} non-scintillating state, decreasing by the same amount the scintillation efficiency. Annealing the crystals under oxygen atmosphere restores the scintillation efficiency by re-oxidizing the Ce^{2+} ions. Ref. [11] provides more details.

The kinetics of the radiation damage build-up and recovery is determined by the depth of the traps at the origin of colour centres. Very shallow traps induce transient absorption bands, which recover so quickly that the monitoring of the crystal transparency becomes very difficult. Much attention has been paid when optimizing PWO crystals for the CMS calorimeter at LHC to suppress as much as possible such defects or to compensate their effect by specific doping [8, 29]. On the other hand, deep traps are generally very stable and are characterized by a continuous increase of the corresponding absorption bands, even at low dose rate, until they are completely saturated. The monitoring of the crystal transparency allows correcting for light yield variations but the concentration of such defects must be maintained small enough to minimize the loss in light yield. For most of the known scintillators a concentration of such defects at the ppm level can produce a radiation induced absorption coefficient limited to about 1 m^{-1}.

At room temperature a large fraction of the radiation induced defects are metastable. Temperature dependant relaxation processes take place in the crystal lattice so that these defects, once produced, are ionized at a rate, which depends on their energy depth and the temperature following the Boltzmann law. As a consequence, the transmission damage reaches a saturation level, which is dose-rate-dependent up to the point where the rate of trapping of the charge carriers induced by radiation is exactly balanced with the rate of spontaneous relaxation at this working temperature. For a uniform distribution of defects of type *i* in the crystal

and in the absence of an interaction between them the kinetics of the concentration of damaged centres of type i is described by the following differential equation:

$$\frac{\mathrm{d}N_i}{\mathrm{d}t} = -\omega_i N_i + \frac{S}{d_i}\left(N_i^* - N_i\right) \tag{3.16}$$

where N_i is the amount of damaged centres of type i at time t, ω_i is their recovery rate, S is the dose rate, N_{i*} is the amount of pre-existing defects of type i and d_i is a damage constant, which depends on the capture cross-section of free carriers by the centres of type i. The induced absorption coefficient μ produced by irradiation is proportional to the concentration of absorbing centres N through $\mu = \sigma N$, where σ is the cross-section of the absorbing centre. The solution of this equation gives the kinetics of the induced absorption build-up:

$$\mu = \mu_{\mathrm{sat}} \frac{S}{S + \omega d}\left\{1 - \exp\left[-\left(\omega + \frac{S}{d}\right)t\right]\right\} \tag{3.17}$$

where $\mu_{sat} = N_* \sigma$ corresponds to the maximum possible saturation when all centres are damaged. The recovery of the transmission after the end of the irradiation at time t_0 is described by:

$$\mu = \mu_{\mathrm{sat}} \frac{S}{S + \omega d}\left\{1 - \exp\left[-\left(\omega + \frac{S}{d}\right)t_0\right]\right\} \exp\left(-\omega\left(t - t_0\right)\right) \tag{3.18}$$

Figure 3.18 illustrates the impact of this behaviour on the light output of a 23 cm long PWO crystal exposed to a cycle of several irradiations separated by periods of recovery.

There are two ways to increase the radiation hardness of scintillating crystals. The first one is to make every effort to reduce the density of point charge defects

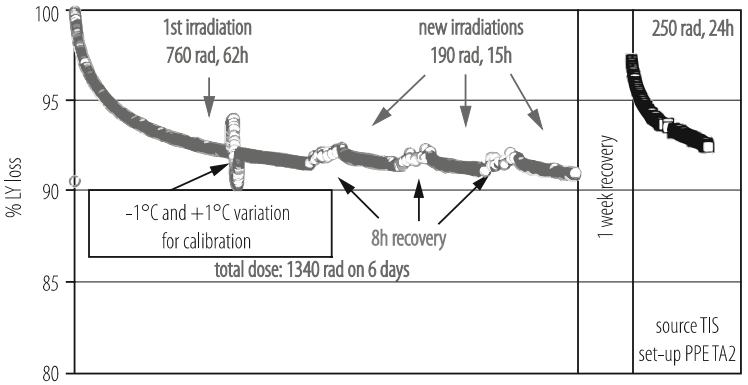


Fig. 3.18 Variations of light out-put for a PWO crystal exposed to a cycle of several irradiations separated by periods of recovery at 18°C (courtesy CMS collaboration)

related to structural defects, impurities and anion or cation vacancies induced by differential evaporation of the chemical components during the crystal growth. This can be achieved for the majority of crystals, through different cycles of purification of the raw materials, multiple crystal growth and annealing of the crystals in specific atmosphere and temperature conditions. This approach is however costly and limited to defect concentration levels in the ppm range. For some applications, such as in high luminosity collider experiments, this is sometimes not enough to guarantee the optical stability of the crystals over long periods.

In another approach additional well selected defects are produced in the crystal, which compete with the uncontrollable defects and reduce their influence. This so-called co-doping strategy has been the result of improved understanding of the mechanisms of light production and charge carrier transport and trapping, opening the way to a defect engineering of the crystals. It has been shown for instance that divalent doping with Ca^{2+} or Mg^{2+} in some Ce^{3+} activated crystals (in particular in ortho-silicates and aluminium garnets), not only increases the light yield, but also suppresses slow scintillation components and improves the radiation hardness [30]. This is the result of easier charge carrier transport to the luminescent centres through the energy levels of these impurities and easier delocalization of trapped carriers due to the smaller energy gap between these traps and the conduction band, which may even absorbed in the conduction band.

3.4 Crystal Engineering. Impact of New Technologies

The conditions of synthesis of the chemical components of a crystal are governed by thermodynamic relations between composition, temperature and pressure of the mixture. At a given pressure, the composition-temperature equilibrium for both the liquid and solid phases is represented by a phase diagram. The phase diagram shows the domains of stability of a given chemical composition and the influence of deviations from stoichiometry (composition of the mixture), unwanted impurities or specific doping. An example of such a phase diagram is shown in Fig. 3.19 for PWO crystals.

Two stable compositions can be grown from a PbO-WO_3 mixture, namely $PbWO_4$ (PWO) and Pb_2WO_5. The $PbWO_4$ melts congruently, i.e. without decomposition of the compound, at 1123°C. The analysis of this phase diagram helps to define some practical parameters for the $PbWO_4$ crystals. First of all the melting temperature restricts the choice of the crucibles to metals with melting points much higher than 1123°C, such as platinum, iridium and their various alloys. Moreover, such crucibles must be chemically inert with melts of similar oxides like $PbMoO_4$, $CaMoO_4$, $ZnWO_4$, as Mo, Ca and Zn are impurities likely to be present in the raw materials. Secondly, the possibility to deviate from the perfect stoichiometric composition of the raw material with some excess of either WO_3, or PbO is of great importance to compensate for a strong differential evaporation of the different components of the melt during the growth process. An initial deviation

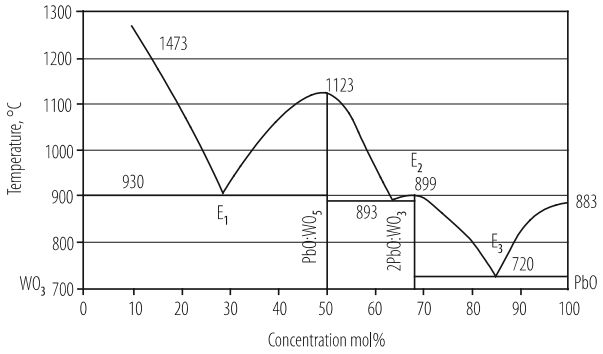


Fig. 3.19 Phase diagram of the PbO-WO_3 system

from the perfect stoichiometry can compensate non-stoichiometry defects. Some restrictions can appear because of segregation processes of additional doping ions. The segregation coefficient k defines how the concentration of doping ions or impurities will vary along the crystal according to the formula:

$$C_s = \frac{kC_0}{1-(1-k)\,g} \tag{3.19}$$

where g is the fraction of the melt already crystallized, C_S is the impurity concentration in the melt at some point, C_0 is the initial impurity concentration in the melt, k is the segregation coefficient. If the segregation coefficient k is too different from 1, as a result of too small or too large ionic radii or different valence states as compared to the ions of the crystal lattice, the doping ion will be pumped in or repelled from the crystal during the growth process.

The majority of crystal growth methods are based on the principle of oriented crystallization. An oriented seed (a small piece of the same crystal or of different composition but similar lattice parameters) is introduced in contact with the melt to initiate the growth process. A temperature gradient is applied so that heat transfer is used as the driving force of crystallization. Several crystallization methods have been developed, which differ in the way the heat transfer and the hydrodynamic conditions are applied:

- Establishing a temperature gradient between the crystal and the melt by heat transfer from the seed. Such heat transfer methods occurred in nature to form crystals and are still used for cheap crystal production, when the requirements on quality are not too high.
- Floating temperature gradient through the melt (Bridgeman and Stockbarger methods). The raw material is placed in a closed crucible, at the end of which a seed has been placed. The crucible is moved through a thermal gradient zone, where the temperature is lowered below the melting point. This is the area where the crystallization takes place. The volume of the melt will therefore decrease

continuously and the growing crystal starts substituting for the melt. This method is relatively inexpensive and multiple crystal pulling is possible by moving several crucibles together through the temperature gradient zone of a single oven [31]. If the simplicity and reliability of the Bridgeman and Stockbarger methods make them particularly attractive for many applications, these methods suffer from several drawbacks, such as large variations of the temperature field parameters during the crystal growth and strong non-uniformities in the distribution of doping ions, impurities and defects in the crystal.

- Establishing a temperature gradient between the crystal and the melt in an open crucible by progressive cooling of the melt after seeding or extracting the growing crystal from melt (Kyropoulos and Czochralski methods, respectively). In the classical Kyropolos method [32] the entire crystallization process starts with the seeding and propagates through the melt as a result of a continuous temperature decrease applied during the process. There is no relative movement of the seed and the crucible. In the Czochralski method [33] the crystal is pulled from the melt. The seed is attached to a Platinum rod and put in contact with the melt in the crucible. The rod or the crucible (sometimes both) are rotating at a few rpm to maintain a good homogeneity of the melt in contact with the crystallized phase. The rod is simultaneously pulled up at a speed of typically 1–10 mm/h depending on the crystal. This method is the most widely used for growing oxide scintillators and several other types of scintillators because of its potential to grow high quality crystals by concentrating impurities and defects in the bottom part of the crucible.

More details about crystal engineering techniques are given in ref. [11].

Technologies for the production of crystals are rapidly evolving. The impressive progress in nanotechnologies in particular open new perspectives for the production of pre-reacted raw materials of excellent quality with a high uniformity of the grain sizes. With these new materials, transparent ceramics of heavy scintillators can be produced (Fig. 3.20), with the advantage over standard crystal growth techniques to be much more cost effective: not only the scintillator can be produced to its final shape, saving on the cost of mechanical processing, but also the temperature for sintering is usually much lower than for standard crystal growth.

The recently developed pulling-down technology from a shape-controlled capillary die gives the possibility to produce elongated crystals with dimensions that are not accessible using traditional cutting and polishing of bulk crystals grown by the more standard Czochralski or Bridgeman methods (Fig. 3.21). This approach has important advantages, such as growing the crystal in the final shape (round, oval, square, rectangular, hexagonal), very rapidly (several millimeters per minute instead of millimeters per hour), simultaneous multifibre pulling, increased activator doping concentration, etc. Excellent quality BGO, YAG and LSO fibers have been grown with a length of up to 2 m and a diameter between 0.3 and 3 mm. Some other materials are being studied, in particular from the very interesting perovskite family: YAP and LuAP [34].

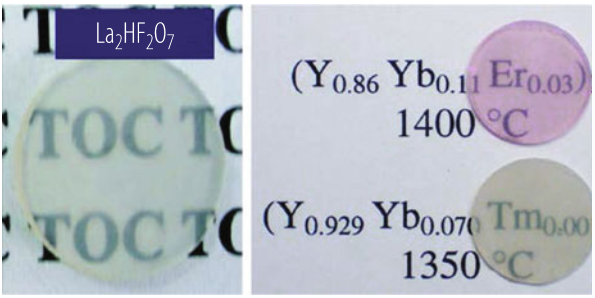


Fig. 3.20 Transparent ceramics of different heavy scintillators prepared with pre-reacted nanopowders

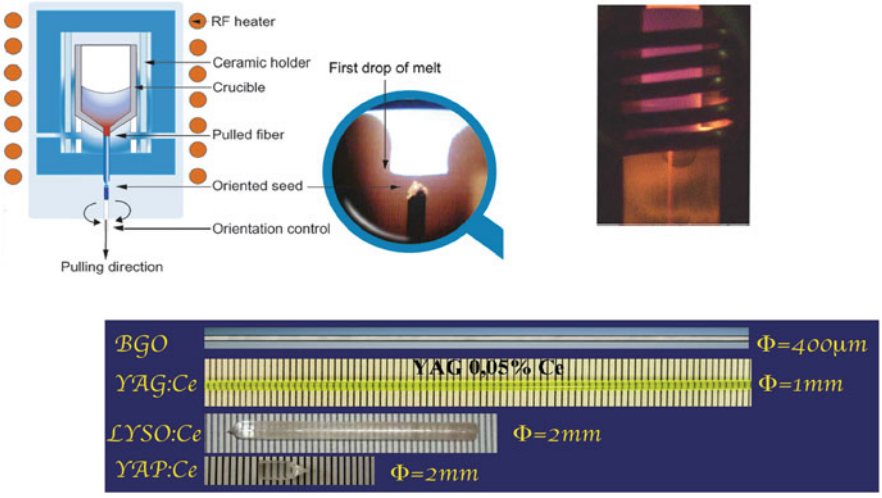


Fig. 3.21 The micro-pulling down crystal growth technology (courtesy Fibercryst)

3.5 Table of Commonly Used Scintillators

Inorganic scintillators generally considered for a majority of applications, and in particular, for particle physics detectors and medical imaging cameras are listed in Table 3.2 with their most important physico-chemical and optical properties. A much more exhaustive list of scintillators classified according to their chemical structure is presented in ref. [11].

Table 3.2 Most commonly used scintillators with their main physico-chemical and optical parameters

Scintillator	Simplified name	Density [g/cm^3]	Light Yield [ph/MeV]	Emission wavelength [nm]	Decay time [ns]	Hygro-scopic	Main application
NaI :Tl		3.67	38,000	415	230	Yes	Medical imaging, industrial γ camera, homeland security
CsI :Tl		4.51	54,000	550	1000	Lightly	Physics detectors
$CdWO_4$	CWO	7.9	28,000	470/540	20,000/5000	No	X-Ray scanner
$(Y,Gd)_2O_3$:Eu	YGO	5.9	19,000	610	1000	No	X-Ray scanner
Gd_2O_2S:Pr,Ce,F	GOS	7.34	21,000	520	3000	No	X-Ray scanner
$Bi_4Ge_3O_{12}$	BGO	7.13	9000	480	300	No	Physics detectors medical imaging
Gd_2SiO_5	GSO	6.7	12,500	440	60	No	Medical imaging
Lu_2SiO_5	LSO	7.4	27,000	420	40	No	Medical imaging
Lu_2AlO_3	LuAP	8.34	10,000	365	17	No	Medical imaging
$LaBr_3$:Ce	BriLanCe™	5.29	61,000	358	35	Very	Medical imaging
BaF_2		4.89	2000/8000	220/310	0.7/620	No	Physics detectors
CeF_3		6.16	2400	310/340	30	No	Physics detectors
$PbWO_4$	PWO	8.28	200	420	5/15	No	Physics detectors

References

1. B. Rossi, High Energy Particles, Prentice-Hall, Inc. Englewood Cliffs, NY, 1952.
2. U. Fano, Annu. Rev. Nucl. Sci. 13 (1963) 1.
3. J.D. Jackson, Classical Electrodynamics, 2nd ed., Wiley, New York, 1975, chap. 13.
4. R.D. Evans, The Atomic Nucleus, Krieger, New York, 1982.
5. A. Lempicki et al., *Fundamental limits of scintillator performance*, Nucl. Instrum. Meth. A 333 (1993) 304-311.
6. P. Lecoq et al., *Lead Tungstate ($PbWO_4$) scintillators for LHC EM calorimetry*, Nucl. Instrum. Meth. A 365 (1995) 291-298.
7. Scintillation Detectors, Catalog, Saint Gobain, Ceramiques Industrielles, March 1992.
8. A. Annenkov et al., *Suppression of the radiation damage in Lead Tungstate scintillation crystal*, Nucl. Instrum. Meth. A 426 (1999) 486-490.
9. The CMS Collaboration (2015) Technical proposal for the phase II upgrade of the compact muon solenoid, CERN-LHCC-2015-010/LHCC-P-008
10. Gundacker S. et al, (2013), Time of Flight positron emission tomography towards 100ps resolution with L(Y)SO: an experimental and theoretical analysis, JINST 8:P07014
11. P. Lecoq, A. Annenkov, A. Gektin, M. Korzhik, C. Pedrini, Inorganic Scintillators for Detector Systems, Springer-Verlag, 2006. Second edition 2017, ISBN 978-3-319-45521-1 doi 10.1007/978-3-319-45522-8
12. A.J. Dean, *Imaging in high-energy astronomy*, in: *Heavy Scintillators for scientific and industrial applications*, Proc. Int. Workshop Cristal 2000, Sep 22-26, 1992, Chamonix, France, F. De Notaristefani, P. Lecoq, M. Schneegans (eds.), Editions Frontières France, 1992, pp. 53-64.
13. C. Kuntner et al., *Intrinsic energy resolution and light output of the Lu0.7Y0.3AP:Ce scintillator*, Nucl. Instrum. Meth. A 493 (2002) 131-136.
14. M.C. Abreu et al., *ClearPEM: A PET imaging system dedicated to breast cancer diagnostics*, Nucl. Instrum. Meth. A 571, (2007) p. 81-84.
15. J.B. Birks, The Theory and Practice of Scintillation Counting, Pergamon, London, 1964.
16. A. Vasiliev, *Relaxation of hot electronic excitations in scintillators: account for scattering, track effects, complicated electronic structure*, Proc. Fifth Int. Conf. Inorganic Scintillators and their Applications, SCINT99, V.V. Mikhailin (ed.), Moscow State University, Moscow, 2000, pp. 43-52.
17. C. Pedrini et al., *Time-resolved luminescence of CeF_3crystals excited by X-ray synchrotron radiation*, Chem. Phys. Lett. 206 (1993) 470–474.
18. M. Kronberger, E. Auffray, P. Lecoq, *Probing the concept of Photonics Crystals on Scintillating Materials*, Proc. 9th Int. Conf. Inorganic Scintillators and their Applications, Winston-Salem, NC, June 4-8, 2007, IEEE Trans. Nucl. Sci. 55(3) (2008) 1102-1106.
19. A. Knapitsch, P. Lecoq, Review on photonic crystal coatings for scintillators, International Journal of Modern Physics A, Vol. 29 (2014) 1430070 (31 pages) https://doi.org/10.1142/S0217751X14300701
20. R.B. Murray, A. Meyer, *Scintillation response of activated inorganic crystals to various charged particles*, Phys. Rev. 122 (1961) 815-826.
21. W.W. Moses, S.A. Payne, W.S. Choong, *Scintillator Non-Proportionality: Present Understanding and Future Challenges*, IEEE Trans. Nucl. Sci. 55(3) (2008) 1049-1053.
22. B.D. Rooney, J.D. Valentine, *Scintillator Light Yield Non-proportionality: Calculating Photon Response Using Measured Electron Response*, IEEE Trans. Nucl. Sci. 44(3) (1997) 509-516.
23. P. Lecoq, M. Korzik, A. Vasiliev, *Can transient phenomena help improving time resolution in scintillators?*, IEEE Trans. Nucl. Sciences, Vol. 61, NO. 1, February 2014
24. J.Q. Grim et al., *Continuous-Wave Biexciton Lasing at Room Temperature Using Solution-Processed Quantum Wells,* Nat. Nanotechnol. **9,** 891–895 (2014)
25. P. Lecoq, *Metamaterials for novel X- or γ-ray detector designs*, 2008 IEEE Nuclear Science Symposium Conference Record, N07-1, pp.680-684

26. A. Annekov, M. Korzhik, P. Lecoq, et al., *Slow components and afterglow in PWO crystal scintillation*, Nucl. Instrum. Meth. A 403 (1998) 302-312.
27. US Patent 5521387
28. Chen Lingyan, Du Jie, Wang Liming, Xiang Kaihua, *An investigation of radiation damage induced by hydroxyl and oxygen impurities in BaF2 crystal*, Scintillator and Phospor Materials MRS Proceeding 348 (1994) 447-454.
29. A. Annenkov, M. Korzhik, P. Lecoq, *Lead tungstate scintillation material*, Nucl. Instrum. Meth. A 490 (2002) 30–50.
30. M. Nikl et al., *Defect Engineering in Ce-Doped Aluminum Garnet Single Crystal Scintillators,* Cryst. Growth Des. 2014, 14, 4827−4833, dx.doi.org/10.1021/cg501005s
31. Z.W. Yin, Q. Deng, D.S. Yan, *Research and Development Works on PbWO4 Crystals in Shanghai Institute of Ceramics*, Proc. 5th Int. Conf. Inorganic Scintillators and their Applications, Moscow, Apr 16-20, 1999, pp. 206-211.
32. S. Kyropolos, F. Zeits, *Ein Verfahren zur Herstellung grosser Kristalle*. Anorg. Allgem. Chem. 154 (1926) 308–313.
33. J. Czochralski, *Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle*, Z. Phys. Chem. 92 (1918) 219-224.
34. B. Hautefeuille et al., J. Crystal Growth 289 (2006) 172.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 4
Gaseous Detectors

H. J. Hilke and W. Riegler

4.1 Introduction

All gaseous detectors signal the passage of charged particles by gathering the electrons from the ion pairs produced in the gas, usually after some amplification. The history of the gas detectors starts with the counter described by Rutherford and Geiger in 1908 [1]. It consisted of a cylindrical metallic tube filled with air or other simple gases at some 5 Torr and with a 0.45 mm diameter wire along its axis. The negative high voltage on the tube with respect to the wire was adjusted to below the discharge limit. With a gas gain of a few 10^3, only α-particles could be detected as current pulses with an electrometer. This counter was the first electronic counter, following the optical counting of light flashes in the study of radioactive substances with scintillating crystals. A major step was taken when Geiger found that by replacing the anode wire by a needle with a fme pin, electrons could also be detected [2]. These *needle counters* became the main particle counter for years. Already in 1924, Greinacher started using electronic tubes to amplify the signals [3].

The *Geiger-Mueller-counter* was first described in 1928 [4]: it produced strong signals independent of the primary ionization. Used with rare gases, these counters required load resistances of $10^8 - 10^9$ Ohm to avoid continuous discharges, resulting in dead times of $10^{-3} - 10^{-4}$ s. Later, external circuits were introduced to shorten the dead time. The real progress, however, brought the discovery in 1935 by Trost [5] that the addition of alcohol quenches the gas discharges internally,

The author H. J. Hilke is deceased at the time of publication.

H. J. Hilke · W. Riegler (✉)
CERN, Geneva, Switzerland
e-mail: Werner.Riegler@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_4

permitting low load resistances and thus high counting rates. Cosmic ray physics in particular profited from systems of such counters used with electron tube coincidence circuits. It took a number of years to understand the basic processes in different gases and under various operation conditions.

Proportional counters regained interest, when the development of more sensitive readout electronics permitted energy determination. In the second half of the 1940s, however, the demand for faster counters with longer lifetime and higher sensitivity initiated a move towards scintillation techniques, which saw a rapid development, especially after the introduction of the photomultiplier, soon providing fast response and time resolutions below 10^{-8} s. On the gas detector side, only the novel technique *of parallel plate counters* [6] could compete, with time resolutions down to 10^{-10} s, however with lower rate capability. A detailed account of the developments up to the 1950s can be found in [7].

The field of gas detectors was revived with the introduction of the *multiwire proportional chamber* by Charpak in 1968 [8] and shortly afterwards with the extension by two groups to *drift chambers* with different geometries [9, 10]. The following decades saw a rapid development of the techniques, especially in high energy physics but also for nuclear physics and other fields. An additional major R&D effort was triggered in the 1990s by the requirements for the LHC: extreme particle rates and radiation hardness. Solutions demanded very careful choice of gas fillings as well as of construction materials and methods. Gas detectors were and are still used mainly for tracking but also in calorimeters, Cherenkov counters and the detection of transition radiation. Only in the layers closest to the interaction points in accelerator experiments and in other applications where spatial resolution is the prime requirement, finely grained silicon detectors have taken over as first choice. Most of the detector developments were made possible only by the extremely rapid progress in the field of electronics, with respect to miniaturization, integration density, cost and radiation hardness.

Powerful simulation programs have been developed in the past decades and have been widely used in the development and optimization of gas detectors. The program *Garfield* [11] calculates electric fields, electron and ion trajectories and induced signals. The program *Heed* [12] describes primary ionization produced by fast particles in gases and the program *Magboltz* [13] electron transport properties in gas mixtures. The agreement of simulation and measurement has become impressive.

We shall at several occasions refer to designs and studies from the LHC experiments. Recent detailed reports them may be found in [14–17]. The development of the last years can well be followed in the Proceedings of the Vienna Conference on Instrumentation initiated in 1977 as Wire Chamber Conference on a tri-annual basis [18] and of the annual IEEE Nuclear Science Symposia.

The following sections will start with a description of the basic processes in gaseous detectors: ionization of the gas by charged particles (Sect. 4.2.1), transport of electrons and ions in electric and magnetic fields (Sect. 4.2.2), avalanche amplification in high electric fields (Sect. 4.2.3), formation of the readout signals (Sect. 4.2.4) and ‘ageing’ of detectors under irradiation (Sect. 4.2.6). A discussion

of major directions of detector design and performance follows in Sect. 4.3: Single-wire tubes (Sect. 4.3.1), Multi-Wire-Proportional Chambers (Sect. 4.3.2), Drift Chambers (Sect. 4.3.3), Resistive Plate Chambers (Sect. 4.3.4) and Micropattern Devices (Sect. 4.3.5).

4.2 Basic Processes

As most processes depend on the velocity of a particle, we shall often state numerical values for minimum ionizing particles (*mip*), i.e. for $\gamma = 3 - 4$.

4.2.1 Gas Ionization by Charged Particles

The passage of charged particles through a gas is signaled by the production of electron/ion pairs along its path. The electrons are attracted by electrodes on positive potential, in the vicinity of which they are usually amplified in a avalanche process. We give a short summary of the various aspects of the ionization processes, following to some extent [19].

4.2.1.1 Primary Clusters

The ionizing collisions of the particle are occurring randomly with a *mean distance* λ, related to the ionization cross-section per electron σ_{I} and the electron density N_{e} of the gas:

$$1/\lambda = N_{\mathrm{e}}\, \sigma_{\mathrm{I}}. \tag{4.1}$$

The number k of ionizing collisions on a path length L thus follows a Poisson distribution with mean L/λ:

$$\mathrm{P}\,(k|L,\ \ \lambda) = \left((L/\lambda)^k/k!\right) \exp\left(-L/\lambda\right). \tag{4.2}$$

The probability to have no ionization in L is

$$\mathrm{P}\,(0|L,\ \ \lambda) = \exp\left(-L/\lambda\right). \tag{4.3}$$

This relation is used to determine λ and defines the *inefficiency* of a counter measuring a track length L, if it is sensitive to a single primary electron.

The probability distribution f(l)dl for a free flight pass l between two ionizing collisions—i.e. the probability of no ionization in l and one in dl-is an exponential,

$$\mathrm{f}(l)\mathrm{d}l = (\mathrm{d}l/\lambda)\exp(-l/\lambda), \tag{4.4}$$

i.e. short distances are favoured.

An electron ejected in a primary collision on atom A may have enough energy to ionize one or more other atoms. Thus *clusters* of two or more electrons are formed by *secondary ionization*. These clusters are mostly rather localized, as the ejection energy is usually low and results in a short range. High ejection energies for so-called δ-*electrons* are rare, their average number per cm is approximately inversely proportional to energy:

$$P(E > E_0) - y/\left(\beta^2 \mathrm{E}_0\right)/\mathrm{cm}, \tag{4.5}$$

with E_0 in keV, and $y = 0.114$ for Ar, and $y = 0.064$ for Ne [20]; β= particle velocity/speed of light in vacuum. Thus, in Ar, for β~1 and $E_0 = 10\,\mathrm{keV}$, $P = 0.011$/cm, i.e., on average one collision with $E > 10\,\mathrm{keV}$ will occur on a track length of 90 cm. The range of a 10 keV electron is about 1.4 mm. The range decreases very rapidly with decreasing energy and is only about 30 μm for a 1 keV electron.

It tums out that, although the majority of the 'clusters' consist of a single electron, clusters of size >1 contribute significantly to the *mean total number* n_T *of electrons produced per cm*, so that n_T is significantly larger than n_p, the *mean number of primary clusters per cm*. Table 4.1 gives experimental values for some of the common detector gases. In Ar at NTP one finds on average $n_\mathrm{p} = 26$ and $n_\mathrm{T} = 100$ electrons/cm for a minimum ionizing particle, where n_T depends on the volume around the track taken into account. The most probable value for the total ionization is $n_\mathrm{mp} = 42$ electrons/cm. The big difference between n_T and n_mp is an indication of the long tail of the distribution.

Table 4.1 Properties of gases at 20°C, 1 atm

Gas	n_p	n_T	w [eV]	E_I [eV]	E_x [eV]	p [mg/cm^3]
He	4.8	7.8	45	24.5	19.8	0.166
Ne	13	50	30	21.6	16.7	0.84
Ar	25	100	26	15.7	11.6	1.66
Xe	41	312	22	12.1	8.4	5.50
CH_4	37	54	30	12.6	8.8	0.67
CO_2	35	100	34	13.8	7.0	1.84
i-butane	90	220	26	10.6	6.5	2.49
CF_4	63	120	54	16.0	10.0	3.78

n_p, n_T mean primary and total number of electron-ion pairs per cm; w: average energy dissipated per ion pair; E_I, E_x: lowest ionization and excitation energy [21]

4.2.1.2 Cluster Size Distribution

The space resolution in gaseous detectors is influenced not only by the Poisson distribution of the primary clusters along the track but also by the *cluster size distribution*, i.e., by the number of electrons in each cluster and their spatial extent. Little was known experimentally (except for some measurements in cloud chambers) until the first detailed theoretical study [22] for Ar at 1 atm and 20°C. Based on the experimental cross-sections for photo absorption, the oscillator strengths and the complex dielectric constants are calculated and from this the distribution of energy transfers larger than the ionization energy (15.7 eV). Finally, a detailed list is obtained for the distribution of cluster sizes for $\gamma = 4$ and $\gamma = 1000$, to estimate the relativistic rise. A cut of 15 keV was applied to the maximum energy transfer, thus concentrating on the local energy deposition. The mean number of clusters is found to be $n_p = 26.6$/cm at $\gamma = 4$, and 35/cm at $\gamma = 1000$. For a MIP, 80.2% of the clusters are found to contain 1 electron, 7.7% two electrons, 2% three electrons, and 1.4% more than 20 electrons.

Several years later, a detailed experimental study of several gases is reported in [23]. For Ar, 66/15/6 and 1.1% of the clusters are found to contain 1/2/3 and $\geq$20 electrons, respectively. The values for low cluster sizes are quite different from the calculated values mentioned above and the calculated bump around 10 electrons is not seen in the measurement, see Fig. 4.1. The authors suggest as a possible explanation that one assumption made in the simulation may not be appropriate, namely that the absorption of virtual photons can be treated like that of real photons, which also leads to the bump at the L-absorption edge. A simpler model starting from measured spectra of electrons ejected in ionizing collisions gives good agreement with the measurements, in particular for the probability of small cluster sizes.

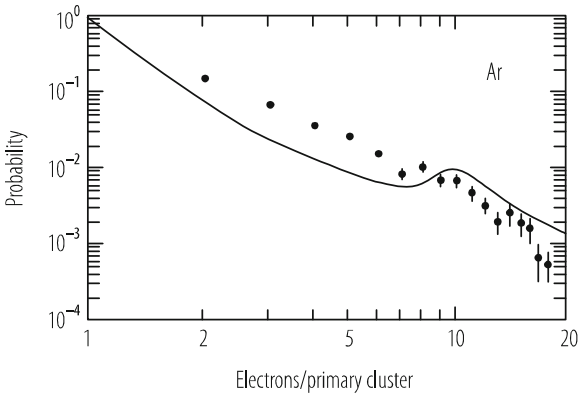


Fig. 4.1 Cluster size distribution: simulation for Ar (continuous line) [22] and measurements in Ar/CH_4 (90/10%) [23]

Space resolution in drift chambers is influenced by the clustering in several ways. The arrival time of the first n electrons, where n times gas amplification is the threshold for the electronics, depends both on the spatial distribution of the clusters and the cluster size. For large clusters, δ-electrons, ionization may extend far off the trajectory.

4.2.1.3 Total Number of Ion Pairs

The detector response is related to the cluster statistics but also to the total ionization n_T, e.g., in energy measurements. A quantity W has been introduced to denote the average energy lost by the ionizing particle for the creation of one ion pair:

$$W = E_i/n_E, \tag{4.6}$$

where E_i is the initial kinetic energy and n_E the average total number of ion pairs *after full dissipation of* E_j.

Measurements of W by total absorption of low energy particles show that it is practically independent of energy above a few keV for electrons and above a few MeV for α-particles. For that reason the differential value w, defined by

$$w = x < \mathrm{d}E/\mathrm{d}x > / < n_T > \tag{4.7}$$

may be used alternatively, as is usually done in Particle Physics, to relate the average total number of ion pairs n_T, created in the track segment of length x, to the average energy lost by the ionizing particle. For relativistic particles, $\mathrm{d}E/\mathrm{d}x$ can not be obtained directly from the difference of initial and final energy (about 270 keV/m for $\gamma = 4$ in Ar), as it is below the measurement resolution. Therefore, w has to be extrapolated from measurements of lower energy particles. For the rare gases one finds $w/I = 1.7 - 1.8$ and for common molecular gases $w/I = 2.1 - 2.5$, where I is the *ionization potential*, indicating the significant fraction of $\mathrm{d}E/\mathrm{d}x$ spent on excitation. Values for photons and electrons are the same, also for α particles in rare gases; in some organic vapours they may be up to 15% higher for α-particles. At low energy, close to the ionization potential, W increases.

In gas mixtures, where an excitation level of component A is higher than I of component B, excited molecules of A often produce a substantial increase in ionization, as has e.g. been observed even with minute impurities in He and Ne: adding 0.13% of Ar to He changed W from 41.3 to 29.7 eV per ion pair. This energy transfer is called *Jesse effect* or *Penning effect*, if metastable states are involved. It is also possible that more than one electron is ejected from a single atom, e.g., by Auger effect following inner shell ionization.

The distribution of n_T in small gas segments is very broad, see an example in Fig. 4.2 [24]. To describe the measurement result, it is thus appropriate to use the *most probable value* instead of the *mean*, since the mean of a small number of measurements will depend strongly on some events from the long tail

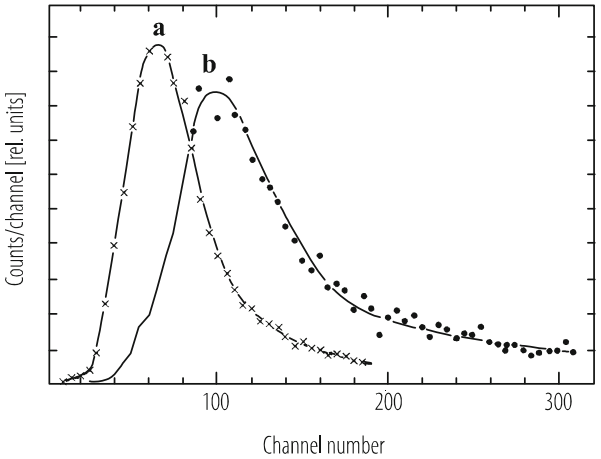


Fig. 4.2 Measured pulse height distribution for 2.3 cm in Ar/CH_4 at 1 atm: (**a**) protons 3 GeV/c, (**b**) electrons 2 GeV/c [24]

of the distribution. The measured pulse height spectrum contains some additional broadening from the fluctuations of the avalanche process. For a mixture of Ar and 5% CH_4, a most probable value of $n_{\mathrm{mp}} = 48$ ion pairs/cm was found for minimum ionizing particles [25].

4.2.1.4 Dependence of Energy Deposit on Particle Velocity

As mentioned above, for position detectors one is interested in the ionization deposited close to the particle trajectory. The Bethe-Bloch formula for dE/dx describes instead the average total energy loss from the incoming particle, including the energy spent on the ejection of energetic δ-electrons which deposit ionization far from the trajectory. To describe the local energy deposit, it is sensible to exclude the contribution from these energetic δ-electrons. This is done by replacing the maximum possible energy transfer T_{max} by a cut-off energy $T_{\mathrm{cut}} << T_{\mathrm{max}}$. This energy cut-off will depend on the experimental conditions and may lie between 30 keV and 1 MeV (in a magnetic field) [19] One then obtains the modified Bethe-Bloch formula for the *mean restricted energy deposit* [20, 21] (see also Chap. 2)

$$\mathrm{d}E/\mathrm{d}x_{\mathrm{restricted}} = K z^2 \, (Z/A) \left(1/\beta^2\right) \left[0.5 \ln \left(2 m_{\mathrm{e}} c^2 \beta^2 \gamma^2 T_{\mathrm{cut}} / I^2\right) - \beta^2/2 - \delta/2\right], \tag{4.8}$$

with $K = 4\pi N_{\mathrm{A}} r_{\mathrm{e}}^2 m_{\mathrm{e}} c^2$, N_{A} = Avogadro constant, m_{e}, r_{e}= mass and classical radius of the electron.

Due to the cut-off, this relation applies not only to heavy particles but also to ionization by electrons [19]. The minimum $\mathrm{d}E/\mathrm{d}x$ deposited by a *minimum ionizing particle* (mip) still lies around $\gamma = 3 - 4$, with $\delta = 0$. For $\beta \to 1$, the *density correction* δ approaches

$$\delta \to 2 \ln \left(h\nu_\mathrm{p} \gamma / I \right) - 1, \tag{4.9}$$

$h\nu_\mathrm{p}$ being the quantum energy of the plasma oscillation of the medium. The restricted energy deposit then reaches a constant value, the *Fermiplateau*, the δ-term compensating the $\ln\gamma$ term:

$$\mathrm{d}E/\mathrm{d}x_\mathrm{restricted} \to \mathrm{P}^2 \, (Z/A) \, 0.5 \ln \left[2mc^2 T_\mathrm{cut} / \left(h\nu_\mathrm{p} \right)^2 \right]. \tag{4.10}$$

In Ar one obtains for the ratio R of energy deposit on the Fermi plateau to the minimum deposit $R = 1.60$, 1.54, and 1.48 for a cut-off $T_\mathrm{cut} = 30$,150 and 1000 keV, respectively [19]. A precise determination of R requires a good estimate of T_cut.

To use the β-dependence of $\mathrm{d}E/\mathrm{d}x$ for particle identification, one has to measure many samples and take their *truncated mean*, e.g., the mean of the lowest 50% pulse heights, to be insensitive to the long tail and to obtain an approximation to the most probable value. See Chap. 2 for details.

4.2.2 Transport of Electrons and Ions

4.2.2.1 Drift Velocities

On the microscopic scale, electrons or ions drifting through a gas are scattered on the gas molecules. In a homogenous electric field E they will acquire a constant *drift velocity* u in the E field direction or, in the presence of an additional magnetic field B, in a direction determined by both fields. Their drift velocities are much smaller than their *instantaneous velocities* c between collisions. Electrons and ions will behave quite differently because of their mass difference.

In the chapters on drift velocities and diffusion we shall follow the argumentation developed in [19]. A relatively simple derivation brings out the main characteristics and does describe a number of experimental results with good approximation. The main approximation of the simple models is to take a single velocity c to represent the motion between collisions. In reality, these velocities c are distributed around a mean value. The shape of the distribution depends on the variation of cross-section and energy loss with the collision velocity. The rigorous theory takes these distributions into account. For lowest velocities there is only elastic scattering, for higher energies various inelastic processes contribute. The elastic and the inelastic spectrum may be described by a single *effective cross-section* $\sigma(c)$ combining the various processes, sometimes called *momentum transfer cross-section*, and by the *average fractional energy loss* $\Delta(c)$ per collision.

Collision cross-sections σ have in some cases been measured directly. Often, however, σ as well as $\Delta(c)$ have to be deduced from measurements of $u(E)$, the dependence of on E, and of diffusion, based on some assumptions on the excitation functions. The consistency of the methods, when applied to other gas mixtures, has improved over the years and is presently very good in a number of practical cases, in particular for the Magboltz simulation [13]; for a comparison of experiments with various models see e.g. [26].

Drift of Electrons

Because of their small mass, electrons will scatter isotropically in a collision and forget any preferential direction. They will acquire a *drift velocity* u given by the product of the acceleration eE/m and the average time τ between collisions

$$u = eE\tau/m. \tag{4.11}$$

Instead of, the notion of *mobility* μ is often used, with μ defined by

$$u = \mu E \rightarrow \mu = e\tau/m. \tag{4.12}$$

Over a drift distance x there will be a balance between the *collision loss* $\Delta\varepsilon_E$ and the energy picked up:

$$(x/u)\,(1/\tau)\,\Delta\varepsilon_E = eEx. \tag{4.13}$$

Here ε_E is the energy gained between collisions, Δ the average fraction of the energy lost in a collision, and $(x/u)(1/\tau)$ the number of collisions on a distance x.

For an *instantaneous velocity* c, the mean time τ between collisions is related to the collision cross- section σ and the number density N of gas molecules by

$$1/\tau = N\sigma c. \tag{4.14}$$

The total energy ε of the electron is given by

$$(m/2)\,c^2 = \epsilon = \epsilon_E + (3/2)\,kT, \tag{4.15}$$

including the thermal energy.

In the approximation $e >> (3/2)kT$, which is often fulfilled for drift of electrons in particle detectors, one obtains

$$\begin{aligned} u^2 &= (eE/mN\sigma)\,\sqrt{(\Delta/2)}\,, \text{ and} \\ c^2 &= (eE/mN\sigma)\,\sqrt{(2/\Delta)} \quad \text{for } \varepsilon - \varepsilon_E >> (3/2)\,kT. \end{aligned} \tag{4.16}$$

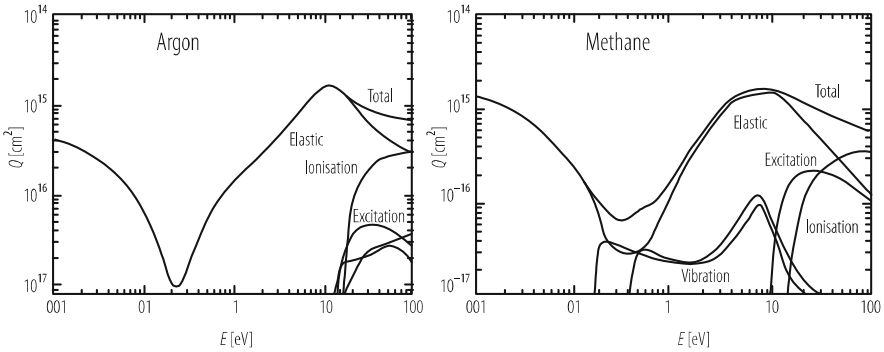


Fig. 4.3 Electron collision cross-sections for Argon and Methane used in Magboltz [13, 27]

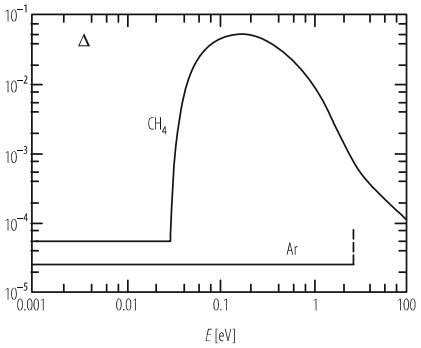


Fig. 4.4 The fraction Δ of energy lost per collision as function of mean energy ε of the electron [28]

The rigorous theory assuming a Dryvestem distribution for the random velocities c adds a multiplication factor of 0.85 to the right sides.

It is important to note that E and N only appear as E/N, the reduced electric field, for which often a special unit is used: one *Townsend* (*Td*) with 1 Td $=10^{-17}$ Vcm^2.

The important role of σ and Δ is obvious; both depend on ε. Below the first excitation level the scattering is elastic and $\Delta - 2m/M - 10^{-4}$ for electrons scattered on gas molecules with mass M. For a high drift speed a small σ is required. Figure 4.3 shows the cross-sections σ for Ar and CH_4. A pronounced minimum, the so-called 'Ramsauer dip' is clearly visible. It leads to high drift velocities in Ar - CH_4 mixtures at low E-values. TPCs take advantage of this.

10^{-17} Vcm^2 =250 V/cm atm at 20°C.

From precise measurements of drift velocity u (to 1%) and longitudinal diffusion D/μ (to 3–5%), σ and Δ have been deduced for some gases [28]. The consistency of the calculated values with measurements of u and D/μ in various other gas *mixtures* gives confidence in the method. Figure 4.4 presents calculated values for Δ as function of ε. Figure 4.5 shows $\varepsilon_k = (2/3)\varepsilon$ derived in the same way in another

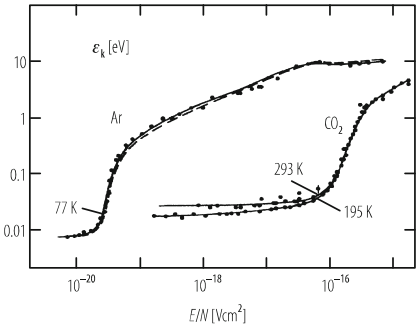


Fig. 4.5 Values for the electron energy ε derived from diffusion measurements as function of the reduced electrical field [29]

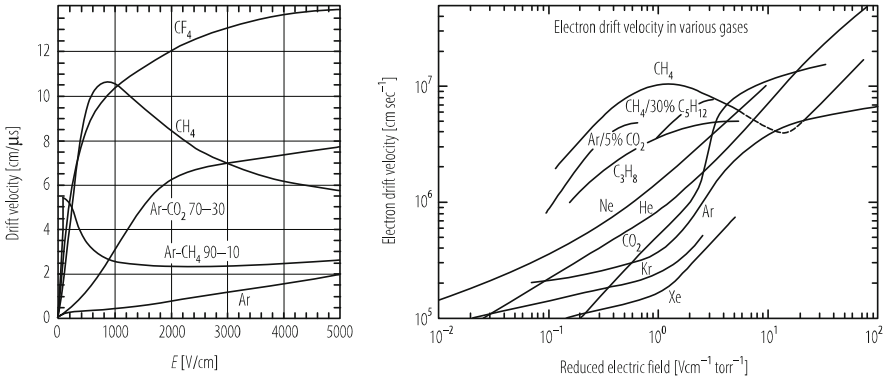


Fig. 4.6 Some examples measured electron drift velocities. (*left*) [30], (*rights*) [31]

study for two extremes, *cold* CO_2 and *hot* Ar [29]. Gases are denoted as *cold*, if ε stays close to the thermal energy $(3/2)kT$ in the fields under consideration. This is the case for gases with vibrational and rotational energy levels, the excitation of which causes inelastic energy losses to the drifting electrons. Cold gases are of interest since they exhibit the smallest possible diffusion.

For *gas mixtures* with number densities $n_i (N = \Sigma\, n_i)$, the effective σ and Δ are given by

$$\sigma = \Sigma n_i \sigma_1 / N, \text{ and}$$

$$\Delta\sigma = \Sigma n_i \Delta_i \sigma_1 / N. \tag{4.17}$$

At low E, drift velocities rise with electric field. Some (e.g. CH_4 and Ar Ar - CH_4 mixtures) go through a maximum, decrease and may rise again. Drift velocities are shown in Fig. 4.6 for some gases [30, 31].

Drift of Ions

Ions of mass m_i acquire the same amount of energy between two collisions as electrons but they lose a large fraction of it in the next collision and their random energy thus remains close to thermal energy. On the other hand the direction of their motion is largely maintained. The result is a much smaller diffusion compared to electrons and constant mobility up to high fields (to ~20 kV/cm atm for A^+ ions in A). In the *approximation for low E field*, the random velocity is considered thermal, i.e. the relative velocity c_rel between the ion and the gas molecules of mass M, which determines τ, is

$$c_\mathrm{rel}^2 = c_\mathrm{ion}^2 + c_\mathrm{gas}^2 = 3kT\left(m_\mathrm{i}^{-1} + M^{-1}\right) \tag{4.18}$$

An argumentation similar to the one followed for electrons [19] leads to

$$u = \left(m_\mathrm{i}^{-1} + M^{-1}\right)^{1/2} (1/3kT)^{1/2} eE/\left(N\sigma\right) \tag{4.19}$$

The ion drift velocity at *low fields* is thus proportional to the electric field. Typical values at 1 atm are around $u = 4$ m/s for $E = 200$ V/cm, to be compared with a thermal velocity around 500 m/s.

In the other extreme of *very high fields*, where thermal motion can be neglected, one finds the drift velocity being proportional to the square root of E. Measurements on noble gas ions [32] in their own gas clearly show both limits with a transition between them at about $15 - 50$ kV/cm atm; see Fig. 4.7. As typical drift fields in drift chambers are a few hundred V/(cm atm), the 'low field approximation' is usually applicable, except in the amplification region.

In a *gas mixture* it is expected that the component with the lowest ionization energy will rapidly become the drifting ion, independently of which atom was ionized in the first place. The charge transfer cross-section is in fact of similar magnitude as the other ion molecule scattering cross-sections. Even impurities rather low concentration might thus participate in the ion migration.

Magnetic Field Effects

A simple macroscopic argumentation introduced by Langevin produces results which are a good approximation in many practical cases.

The motion of a charged particle is described by

$$m\mathrm{d}\mathbf{u}/\mathrm{d}t = e\mathbf{E} + e\,[\mathbf{u}\ \mathrm{x}\ \mathbf{B}] - k\ \mathbf{u}, \tag{4.20}$$

where m, e and $\mathbf{u}$ are the particle's mass, charge and velocity vector, respectively; $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic field vectors; k describes a frictional force proportional to $-\mathbf{u}$.

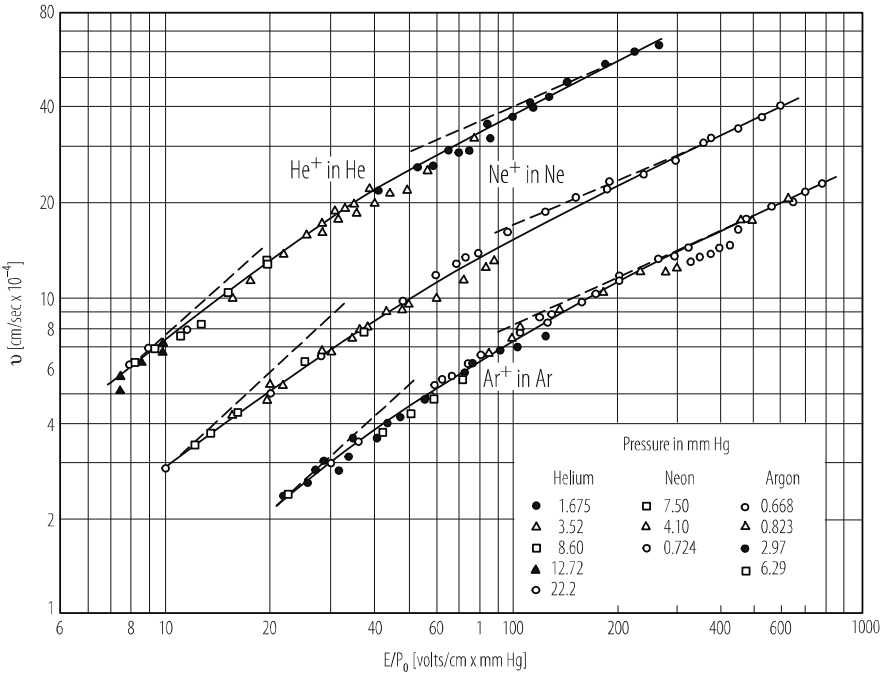


Fig. 4.7 Drift velocities of singly charged ions of noble gases [32]

In the steady state d**u**/dt = 0 and

$$\mathbf{u}/\tau - (e/m)\,[\mathbf{u}\ \mathrm{x}\ \mathbf{B}] = (e/m)\,\mathbf{E}, \tag{4.21}$$

with τ = m/k. The solution for **u** is

$$\mathbf{u} = (e/m)\,\tau \mid \mathbf{E} \mid \left(1/\left(1+\omega^2\tau^2\right)\right)\left\{\mathbf{E}^* + \omega\tau\left[\mathbf{E}^*\ \mathrm{x}\ \mathbf{B}^*\right] + \omega^2\tau^2\,(\mathbf{E}^*\mathbf{B}^*)\,\mathbf{B}^*\right\}, \tag{4.22}$$

where $\omega = (e/m) \mid \mathbf{B}\mid$, and ω carries the sign of e and $\mathbf{E}^*$ and $\mathbf{B}^*$ are unit vectors.

For ions, $\omega\tau \approx 10^{-10}$ Therefore, magnetic fields have negligible effect on ion drift.

For electrons, **u** is along **E**, if $B = 0$, with

$$\mathbf{u} = (e/m)\,\tau\mathbf{E}. \tag{4.23}$$

This is the same relation as the one derived from the microscopic picture (4.11), which provides the interpretation of τ as the *mean time between collisions*.

For large $\omega\tau$, **u** tends to be along **B**, but if **EB** = 0, large $\omega\tau$ tums u in the direction of **ExB**.

Two *special cases* are of practical interest for *electron drift*:

E orthogonal to B

With $\mathbf{EB} = 0$ and choosing $\mathbf{E} = (E_x,\ 0, 0)$ and $\mathbf{B} = (\mathrm{O},\ 0,\ B_z)$, we get

$$\begin{aligned} u_x &= (e/m)\,\tau \mid \mathbf{E} \mid / \left(1 + \omega^2\tau^2\right)\ , \\ u_y &= -\,(e/m)\,\tau\omega\tau \mid \mathbf{E} \mid / \left(1 + \omega^2\tau^2\right), \\ u_z &= 0, \end{aligned} \tag{4.24}$$

and

$$\mathrm{tg}\psi = u_y/u_x = -\omega\tau. \tag{4.25}$$

The latter relation is used to determine $\omega\tau$, i.e. τ, from a measurement of the drift angle ψ, the so-called *Lorentz angle*. In detectors, this angle increases the spread of arrival times and sometimes also the spatial spread. A small $\omega\tau$ would, therefore, be an advantage but good momentum resolution requires usually a strong B.

The absolute value of $\mathbf{u}$ is

$$\mid \mathbf{u} \mid = (e/m)\,\tau \mid \mathbf{E} \mid \left(1 + \omega^2\tau^2\right)^{-1/2} = (e/m)\,\tau \mid \mathbf{E} \mid \cos\psi. \tag{4.26}$$

This means that, independent of the drift direction, the component of $\mathbf{E}$ along $\mathbf{u}$ determines the drift velocity (Tonks' theorem). This is well verified experimentally.

E Nearly Parallel to B

This is the case in the Time Projection Chamber (TPC). Assuming E along z and the components $\mid B_X \mid$ and $\mid B_y \mid \ < < \ \mid B_z \mid$, one finds in first order

$$\begin{aligned} u_x/u_z &= \left(-\omega\tau\, B_y/B_z + \omega^2\tau^2 \mathrm{B_x/B_z}\right) / \left(1 + \omega^2\tau^2\right), \text{ and} \\ u_y/u_z &= \left(\omega x \mathrm{B_x/B_z} + \omega^2\tau^2 \mathrm{B_y/B_z}\right) / \left(1 + \omega^2\tau^2\right). \end{aligned} \tag{4.27}$$

In a TPC this will produce a displacement after a drift length L of $\delta_x = Lu_x/u_z$ and $\delta_y = Lu_y/u_z$ From measurements with both field polarities and different fields, B_X, B_y and τ can be determined.

If B_x and B_y can be neglected with respect to B_z, u_z remains unaffected by B.

4.2.2.2 Diffusion

Due to the random nature of the collisions, the individual drift velocity of an electron or ion deviates from the average. In the simplest case of isotropic deviations, a point-like cloud starting its drift at $t = 0$ from the origin in the z direction will at time t assume a Gaussian density distribution

$$N = (4\pi Dt)^{-3/2} \exp\left(-r^2/(4Dt)\right), \tag{4.28}$$

with $r^2 = x^2 + y^2 + (z - ut)^2$, D being the *diffusion coefficient*. In any direction from the cloud centre, the mean squared deviation of the electrons is

$$\sigma_{\mathrm{I}} = (2Dt)^{1/2} = (2Dz/u)^{1/2} = D^* z^{1/2}. \tag{4.29}$$

with D^* called *diffusion constant*. In terms of the microscopic picture, D is given by

$$D = \lambda^2/(3\tau) = c\lambda/3 = c^2\tau/3 = (2/3)(\varepsilon/m)\tau, \tag{4.30}$$

with λ being the *mean free path*, $\lambda = c\tau$, and ε the *mean energy*.

With the *mobility* μ defined by

$$\mu = (e/m)\tau, \tag{4.31}$$

the mean energy ε can be determined by a measurement of the ratio D/μ:

$$\varepsilon = (3/2)(D/\mu)e. \tag{4.32}$$

Instead of ε, the *characteristic energy* $\varepsilon_k = (2/3)\varepsilon$ is often used.

The *diffusion width* σ_x of an initially point-like electron cloud having drifted a distance L is determined by the electron energy ε:

$$\sigma_X^2 = 2Dt \quad = 2DL/(\mu E) = (4/3)\varepsilon L/(eE) \tag{4.33}$$

This relation is used for the *determination* of D *and* ε.

For a good spatial resolution in drift chambers, a low electron energy and high electric fields are required. The lower limit for ε is the thermal energy $\varepsilon_{\mathrm{th}} = (3/2)kT$. In this limit, the relationship known as *Einstein or Nernst-Townsend formula* follows:

$$D/\mu = kT/e. \tag{4.34}$$

The *minimum diffusion width* is thus

$$\sigma_{x,\mathrm{mm}}{}^2 = (kT/e)(2L/E). \tag{4.35}$$

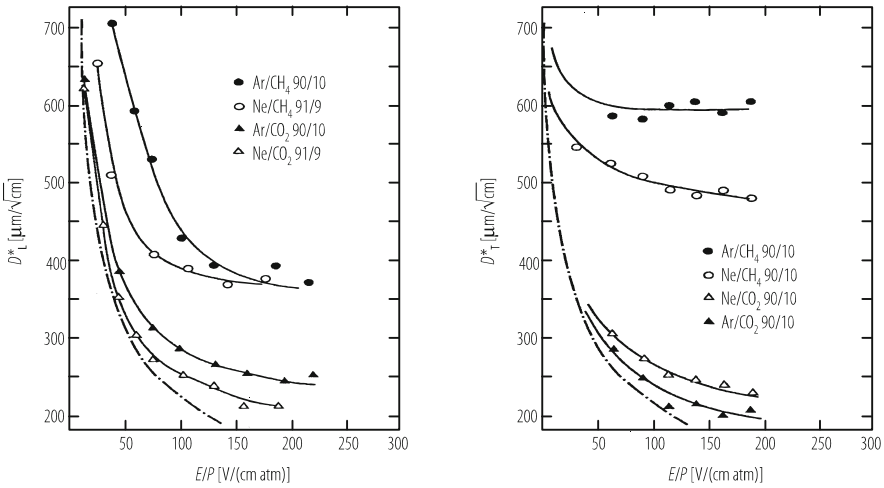


Fig. 4.8 Longitudinal and transverse diffusion constants for low electric fields [33]. The dash-dotted line denotes the thermal limit

As can be seen in Fig. 4.8, this minimum is approached for 'cold gases' like Ar/CO_2 up to E~150 V/cm at 1 atm, for 'hot gases' like Ar/CH_4 only for much lower fields.

Anisotropic Diffusion

So far, we have assumed isotropic diffusion. In 1967 it was found experimentally [34], that the *longitudinal diffusion* D_L along E can be different from the *transversal diffusion* D_T Subsequently it has been established that this is usually the case.

For ions this anisotropy occurs only at high E. As in a collision ions retain their direction to a large extent, the instantaneous velocity has a preferential direction along E. This causes diffusion to be larger longitudinally. However, this high field region is beyond the drift fields used in practical detectors.

For electrons a semi-quantitative treatment [35], restricted to energy loss by elastic collisions, shows that

$$D_L/D_T = (1+\gamma)\,/\,(1+2\gamma) \;\text{ with }\; \gamma = (\varepsilon_0/v_0)\,(\delta v/\delta\epsilon)\,. \tag{4.36}$$

It follows that longitudinal and transversal diffusion will be different, if the collision rate v depends on the electron energy ε.

Figures 4.8, 4.9, 4.10 show measured diffusion for a drift of 1 cm for some common gas mixtures [30, 33, 36]. Simulated diffusion curves are compiled in [33].

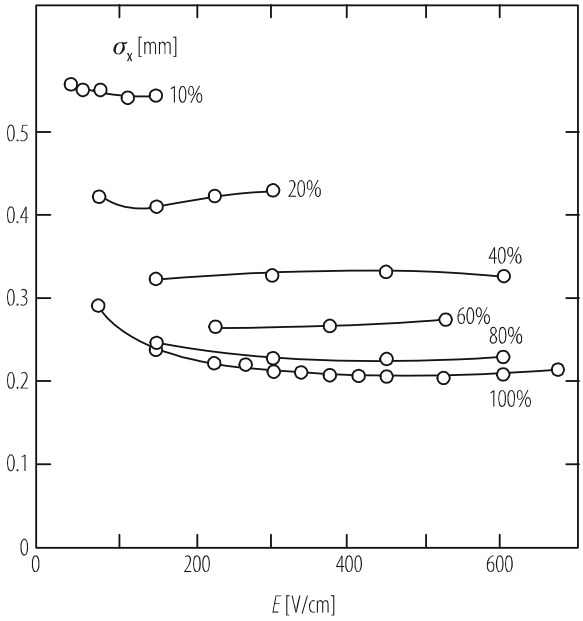


Fig. 4.9 Transverse diffusion for 1 cm drift in Ar/CH_4 mixtures; CH_4 % is indicated [36].

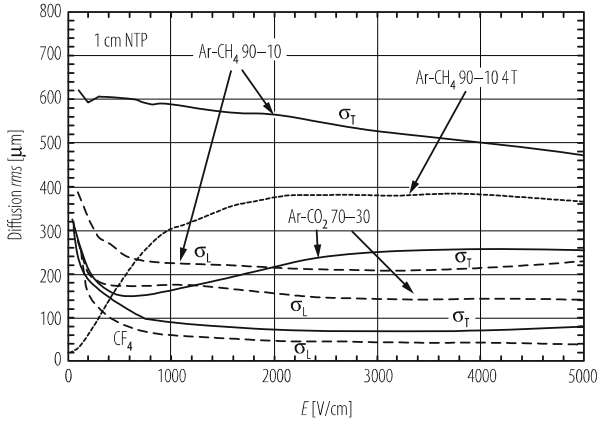


Fig. 4.10 Transverse and longitudinal diffusion for 1 cm drift up to higher E fields [30]

A *magnetic field B* along z will cause electrons to move in circles in the x-y projections in between collisions. The random propagation is diminished and the transverse diffusion will be reduced:

$$D_{\mathrm{T}}(\omega)/D_{\mathrm{T}}(0) = 1/\left(1+\omega^2\tau^2\right). \tag{4.37}$$

This reduction is essential for most TPCs with their long drift distances.

A more rigorous treatment of averages [19] shows that different ratios apply to low and high B:

$$\begin{aligned} D(0)/D(B) &= 1 + \omega^2\tau_1^2 \quad \text{for low } B, \text{ and} \\ D(0)/D(B) &= C + \omega^2\tau_2^2 \quad \text{for high } B. \end{aligned} \tag{4.38}$$

This behaviour was indeed verified [37], by measuring $D(B)$ over a wide range of B. In an Ar/CH_4 (91/9%) mixture the data could be fitted with $\tau_1 = 40$ ps, $\tau_2 = 27$ ps and $C = 2.8$. The high field behaviour is approached above about 3 kg, close to $\omega\tau = 1$.

The longitudinal diffusion remains unchanged: $D_L(\omega) = D_L(0)$.

The effects of E and B combine if both fields are present.

4.2.2.3 Electron Attachment

In the presence of electronegative components or impurities in the gas mixture, the drifting electrons may be absorbed by the formation of negative ions. Halogenides (e.g. CF_4) and oxygen have particularly strong electron affinities. Two-body and three-body attachment processes are distinguished [38].

In the *two-body process*, the molecule may or may not be broken up:

$$\begin{aligned} &e^- + AX \rightarrow Ax^- * \rightarrow A \ (\text{or } A^*) + X^- \left(\text{or } X^- *\right) \text{ or} \\ &e^- + AX \rightarrow Ax^- * \rightarrow AX^- + \text{energy}. \end{aligned} \tag{4.39}$$

The attachment rate R is proportional to the density N:

$$R = c\sigma N \tag{4.40}$$

for an electron velocity c and attachment cross-section σ The rate constants of freons and many other halogen-containing compounds are known [39].

The best known three-body process is the Bloch-Bradbury process [40]. In this process, an electron is attached to a molecule through the stabilizing action of another molecule. It is important for the attachment of electrons with energy below 1 eV to O_2, forming an excited unstable state with a lifetime τ of the order of 10^{-10} s. A stable ion will be formed only if the excitation energy is carried away during τ by another molecule. The attachment rate is proportional to the square of the gas pressure, as it depends on the product of the concentrations of oxygen and of the stabilizing molecules [19]:

$$R = \tau c_e c_2 \sigma_1 \sigma_2 N\,(O_2)\, N\,(X)\,. \tag{4.41}$$

Here c_e is the electron velocity, c_2 the relative thermal velocity between O_2 and X In an Ar/CH_4 (80/20%) mixture at 8.5 atm with an O_2 contamination of 1 ppm, an absorption of 3%/m was measured at a drift speed of 6 cm/μs.

4.2.3 Avalanche Amplification

4.2.3.1 Operation Modes

Gas detectors generally use gas amplification in the homogeneous field of a parallel plate geometry or, more frequently, in the inhomogeneous field around a thin wire. We shall start with the discussion of the second case.

Near a wire with a charge q_s per cm, the electric field at a distance r from its centre is

$$E = q_s / (2\pi\varepsilon_0 r) . \tag{4.42}$$

When raising the field beyond the *ionization chamber regime*, in which all primary charges are collected *without* any *amplification*, at some distance from the wire a field is reached, in which an electron can gain enough energy to ionize the gas and to start an avalanche. The avalanche will grow until all electrons have arrived on the anode wire. For a *gas amplification A* of $1000 \sim 2^{10}$, some 10 ionization generations are required. As the mean free path between collisions is of the order of microns, the field to start an avalanche has to be several times 10^4 V/cm. This is usually achieved by applying a voltage of a few kV to a thin wire, with a diameter in the $20 - 50$ μm range.

Besides ionization, *excitation* will always occur and with it photon emission. A fraction of these photons may be energetic enough to produce further ionization in the gas or on the cathode. Only those photons which ionize outside the radius r_{av} of the moving electron avalanche may be harmful, as their avalanches will arrive later. If γ called the *second Townsend coefficient*, is the probability per ion pair in the first avalanche to produce one new electron, and if A denotes the amplification of the first avalanche, *breakdown* will occur for

$$A\gamma > 1. \tag{4.43}$$

In this case the first avalanche will be followed by a bigger one, this by an even bigger and so on, until the current is limited by external means. If $A\gamma << 1$, $A\gamma$ gives the probability for producing an after-discharge. If a photoionization takes place inside r_{av}, the effect will be an increase of A.

The resulting need to suppress far-traveling photons produced in the rare gases is the reason for the use of 'quench gases' like Methane, Ethane, CO_2, etc., which have large absorption coefficients for UV photons.

The positive ions produced in the avalanche have too little energy to contribute to the ionization in the avalanche. They will move slowly to the cathode(s), where they get neutralized but where rare gas ions may also liberate additional electrons. The addition of the quencher reduces this risk significantly, as its recombination energy can be dissipated in other ways, e.g. by disintegration. This explains why more complex molecules provide higher protection.

Up to a certain value A_p, one has a *proportional regime*: the signal produced will on average be proportional to the number of primary electrons. The amplification will rise approximately exponentially with voltage. The azimuthal extension of the avalanche around the wire will grow with amplification and eventually the avalanche will surround the wire.

When the field is raised above this proportional regime, *space charge effects* will set in. The space charge of the positive ions—moving only very slowly compared to the electrons—will reduce the field at the head of the avalanche and the amplification will rise more slowly with voltage and will no longer be proportional to the primary ionization. In addition, space charge effects will depend on the track angle with respect to the wire and on the density of the primary ionization. This is the so-called *limited proportionality regime*.

Increasing the field further, the positive space charge may produce additional effects. Near the avalanche tail the electric field is increased. If the absorption of UV photons in the quench gas is high, the photons may ionize this high field region and start a *limited streamer* moving backwards by starting avalanches further and further away from the sense wire. As the electric field at large radius weakens, this development will stop after typically 1–3 mm. The total charge is almost independent of the primary charge starting the streamer. The process depends quite strongly on experimental conditions. An example is presented in Fig. 4.11, which shows a steep step from the proportional regime [41]. In the narrow transition zone one finds a rapid change of the ratio of streamer/proportional signal rates. In other experimental conditions a smoother transition has been observed.

If the absorption of the UV photons is weak, photons travel further and avalanches may be started over the full length of the wire, leading to the *Geiger mode*, if the discharge is limited by external means.

4.2.3.2 Gas Gain

With multiplication, the number n of electrons will grow on a path ds by

$$\mathrm{d}n = n\ \alpha\ \mathrm{d}s, \tag{4.44}$$

where α is the *first Townsend coefficient*. Ionization growth is obviously proportional to the gas density p and depends on the ionization cross-sections, which are a function of the instantaneous energy ε of the electrons. This energy itself is a function of E/p. The relationship between α and E is, therefore, given in the form α/p as function of E/p or for a specific temperature as $\alpha/p(E/p)$. Figure 4.12 gives some

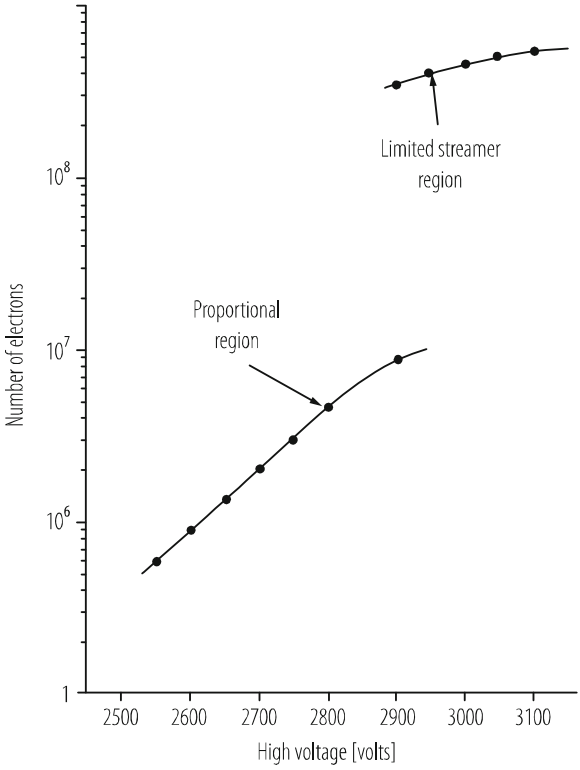


Fig. 4.11 Pulse-height transition from limited proportionality to limited streamer mode [41]

examples of measurements [42]; it shows the strong increase of α with electric field in the region of interest to gas detectors, up to about 250 kV/cm. No simple relation exists for α as function of electric field E, but Monte Carlo simulation has been used to evaluate α. Figure 4.13 shows an example [27]. For the lower field values there is reasonable agreement with measurement. The discrepancy at the highest fields is attributed to photo-and Penning-ionization not being included.

The amplification A in the detector is obtained by integration

$$A = n/n_0 = \exp \int \alpha(s)\, \mathrm{d}s = \exp \int \alpha(E)/\left(\mathrm{d}E/\mathrm{d}s\right) \mathrm{d}E, \tag{4.45}$$

from $E_{\min}$, the minimum field to start the avalanche, to the field $E(a)$ on the wire. $E_{\min} e$ is equal to the ionization energy of the gas molecules divided by the mean free path between collisions. Near the wire and far from other electrodes, the field is

$$E(r) = q_{\mathrm{s}}/\left(2\pi r \varepsilon_0\right), \tag{4.46}$$

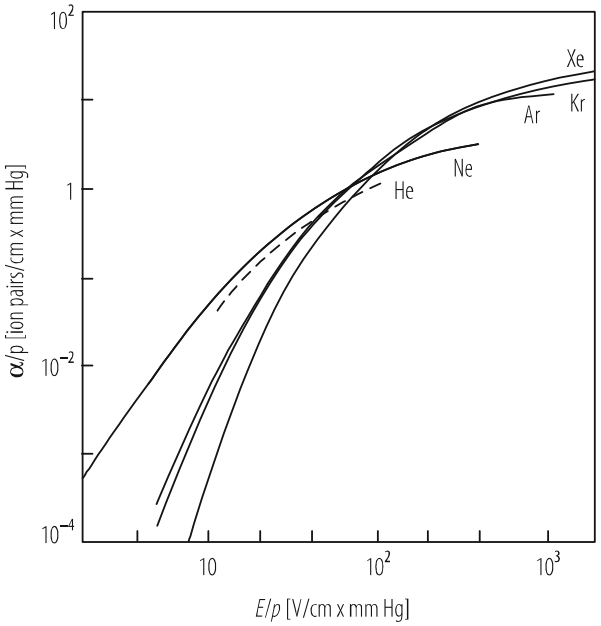


Fig. 4.12 Examples of measured 'First Townsend coefficient' α in rare gases [42]

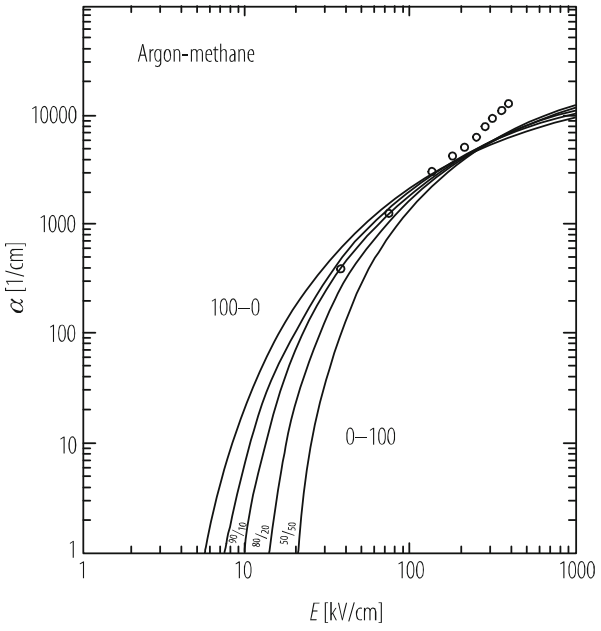


Fig. 4.13 Simulated 'First Townsend coefficient' α in Ar/CH_4 mixtures at 1 atm (100–0 means 100%Ar). Measured values are indicated as circles [27]

where q_s is the charge per cm. Therefore,

$$A = \exp \int q_s \, \alpha(E) \, \mathrm{d}E / \left(2\pi \varepsilon_0 E^2\right). \tag{4.47}$$

Two approximations in particular have been used to describe practical cases. The early *Korff model* [43] uses the parameterization

$$\alpha / p = A \exp\left(-Bp/E\right), \tag{4.48}$$

with empirical constants A and B depending on the gas.

In the *Diethorn approximation* [44], α is assumed to be proportional to E. One then obtains for a proportional tube with wire radius a and tube radius b

$$\ln A = \left(\ln 2 / \ln\left(b/a\right)\right) \left(V/\Delta V\right) \ln\left(V / \left(\ln\left(b/a\right) a E_{\min}\right)\right), \tag{4.49}$$

where the two parameters E_{mm} and ΔV are obtained from measurements of$A$ at various voltages and gas pressures. E_{mm} is the minimum E field to start the avalanche and $e\Delta V$ the average energy required to produce one more electron. E_{mm} is defined for a density p_0 at STP. For another density $E_{\mathrm{mm}}(p) = E_{\mathrm{mm}}(p_0)(p/p_0)$. A list for E_{mm} and ΔV for various gases is given, e.g., in [19]. Reasonable agreement with the experimental data is obtained; discrepancies show up at high A.

4.2.3.3 Dependence of Amplification on Various Factors

The gas amplification depends on many operational and geometrical parameters. Some examples are:

Gas Density

The Diethom approximation gives

$$\mathrm{d}A/A = -\left(\ln 2 / \ln\left(b/a\right)\right) \left(V/\Delta\nabla\right) \left(\mathrm{d}p/p\right) \rightarrow = \text{(5-8)} \;\; \mathrm{d}p/p \tag{4.50}$$

typically.

Geometrical Imperfections

The effects will obviously depend on the geometry and the operation details. An early publication [45] gives analytic estimates of the effects of wire displacements and variations in wire diameter. In a typical geometry $\mathrm{d}A/A \sim 2.5\, \mathrm{d}r/r$, where r is the wire radius; $\mathrm{d}A/A \sim 9 \Delta gap/gap$.

Edge Effects

Near edges, the electric field is reduced over distances similar to the gap between the electrode planes. It can be recovered largely by additional field shaping lines on the edges [46].

Space Charge

Due to the low velocity of the positive ions (falling off as $1/r$ from >1mm/μs at $r = a$), space charge will build up at high particle fluxes and lower the avalanche amplification. In drift tubes the voltage drop due to the space charge from a given particle flux is proportional to the third power of the tube radius. A smaller radius thus improves the rate capability drastically.

4.2.3.4 Statistical Fluctuations of the Amplification

In the proportional regime, the amplification A is simply defined by $A = n/n_{\mathrm{T}}$ and one assumes that each of the n_{T} initial electrons produces on average the same A ion pairs. We define $P(n)$ as the probability to produce n electrons in the individual avalanche with mean A and variance σ^2 If $n_{\mathrm{T}} > > 1$ and if all avalanches develop independently, it follows from the central limit theorem that the distribution function $F(n)$ for the sum of the n_{T} avalanches approaches a Gaussian with mean $n = n_{\mathrm{T}}A$ and variance $S^2 = n_{\mathrm{T}}\sigma^2$, independent of the actual $P(n)$.

On the other hand, for detection of single or a few electrons, knowledge of the individual $P(n)$ is required.

For a parallel plate geometry calculations [47] agree well with measurements [48]. The distributions found theoretically [49] and experimentally [50] for the strong inhomogeneous field around a thin wire also look similar and approach Polya distributions (Fig. 4.14).

For these distributions

$$(\sigma_A / < A >)^2 = f, \text{with } f \leq 1. \tag{4.51}$$

The limiting case $f = 1$ is an exponential distribution (Yule-Furry law)

$$P(A) = (1/ < A >) \exp(-A/ < A >). \tag{4.52}$$

Experimental results point to $f = 0.6 - 1.0$. Measurements with laser tracks [19], indicate that the r.m.s. width σ_A of a single-electron avalanche is close to the mean, as it is for the Polya distribution with $f = 1$. An approximately exponential distribution for single-electron avalanches is also reported in [28, 48].

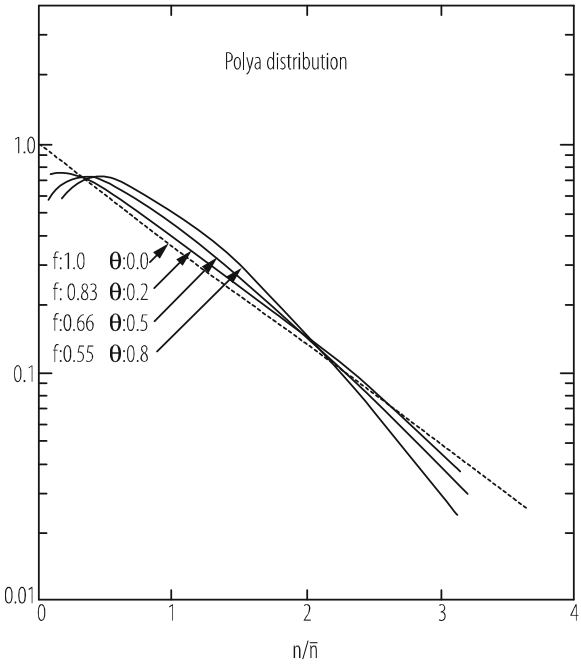


Fig. 4.14 Polya distributions [22]

4.2.4 Signal Formation

In wire chambers, signal formation is very similar to the one in the simplest geometry of a cylindrical tube with a coaxial wire, because most of the useful signal is produced in the immediate vicinity of the sense wire and the electric field around the sense wires in a MWPC can be considered as radial up to a radius equal to about one tenth of the distance between sense wires [45].

Signals are *always* produced *by induction* from the moving charges.

Ramo [51] and Shockley [52] have shown that in general the current I_R induced on the readout electrode R is given by

$$I_R = -q\ \mathbf{E}_w \mathbf{v}, \tag{4.53}$$

where q is the signed charge moving with the vectorial velocity $\mathbf{v}$, and $\mathbf{E}_w$ is a vectorial *weighting field*, a conceptual field defined by applying + 1V on R and 0 V on all other electrodes. The unit of E_w is 1/cm. The actual v is calculated by applying the normal operation voltages, including possibly a B field.

In the special case of a two electrode system like the wire tube, $\mathbf{E}_w = \mathbf{E}_{op}/V$, where $\mathbf{E}_{op}$ is the actual operating field obtained with the voltage V on R (the anode wire) and zero V on the cathode.

For the proportional tube with wire radius a and cathode radius b, E_{op} and E_{w} are obviously radial with

$$E_{\mathrm{op}} = V/\left[r \ln\left(b/a\right)\right]. \tag{4.54}$$

We assume constant mobility μ for the positive ions. Therefore

$$v^{+}\left(\mathrm{t}\right) = \mu V/\left[r(t) \ln\left(b/a\right)\right]. \tag{4.55}$$

For an ion starting at $t = 0$ from $r = r_1$,

$$r(t) = r_1(1 + (t/t_0))^{1/2} \text{ with } t_0 = r_1^2 \ln\left(b/a\right)/\left(2\mu V\right). \tag{4.56}$$

The maximum time for an ion to drift from a to b is

$$T^{+}{}_{\max} = (b/a)^2 t_0, \text{ as } (b/a)^2 >> 1. \tag{4.57}$$

The induced current I^{+} is

$$I^{+} = -q\ \mathbf{E}_{\mathrm{w}}\mathbf{v}^{+} < 0, \tag{4.58}$$

as $\mathbf{v}^{+}$ is parallel to $\mathbf{E}_{\mathrm{w}}$.

For the integrated charge Q, one obtains

$$Q^{+}(t) = \int I\ \mathrm{d}t = \int I\left(1/v^{+}\right)\ \mathrm{d}r = \int -q\,E_{\mathrm{w}}\ \mathrm{d}r. \tag{4.59}$$

Integration from r_1 to r_2 gives

$$Q^{+}{}_{1\to 2} = -q \ln\left(r_2/r_1\right)/\ln\left(b/a\right), \text{ with } q > 0 \text{ and } r_2 > r_1, \tag{4.60}$$

For an electron one obtains

$$Q^{-}{}_{1\to 2} = +q\ |\ln\left(r_2/r_1\right)|\ /\ln\left(b/a\right), \text{ with } q < 0 \text{ and } r_2 < r_1, \tag{4.61}$$

as $\mathbf{v}$ is antiparallel to $\mathbf{E}_{\mathrm{w}}$.

We shall give numbers for a typical proportional tube with $a = 10\ \mu\mathrm{m}$, $b = 2.5$ mm, $E_{\mathrm{op}}(r = a) = 200$ kV/cm, $\mu^{+} = 1.9$ atm cm^2/(Vs), $\mathrm{v}^{-} \approx 5 \cdot 10^6$ cm/s and—to estimate the gas amplification A—the Diethom parametrization $\alpha = (\ln 2/\Delta\nabla)E$ and $E_{\min} = V/(r_{\mathrm{mm}} \ln\left(b/a\right))$, taking for an Ar/$CH_4$(90/10) mixture $\Delta V = 23.6$ V and $\mathrm{E}_{\min} = 48$ kV/cm [19], p. 136. Here r_{mm} is the starting radius for the avalanche and $E_{\min}$ the minimum field permitting multiplication. We obtain: $t_0 = 1.3$ ns, $T^{+}_{\max} = 82\ \mu\mathrm{s}$, $r_{\min} = 42\ \mu\mathrm{m}$, $A = 4400$.

The last electron will be collected in a very short time of about 0.6 ns, the vast majority even faster. Half of the electrons move only about 2 μm, the next 25%

some 4 μm and so on. A rough estimate of the induced electron charge signal is, therefore,

$$Q^{-}{}_{\text{total}} = q \ln(14/10) / \ln(2500/10) = 0.06\, q. \tag{4.62}$$

Only about 6% of the total induced signal is due to the movement of the electrons, the rest from the ions, if one integrates over the full ion collection time.

In practice, however, one mostly uses much faster integration. The long tail in the signal caused by the very slow ion movement has to be corrected for by electronic pulse shaping to avoid pile-up at high rates (see Sect. 4.69). If one uses fast pulse shaping, say 20 ns integration, only a fraction of the ion charge will be seen: an ion starting at $r_1 = a$, reaches $r_2 = 40\,\mu\text{m}$ in 20 ns and induces about 25% of its charge. That means: with 20 ns pulse shaping, one may expect to see an *effective charge* of about 30% of the total charge produced, of which one fifth is due to the electrons.

4.2.5 Limits to Space Resolution

The space resolution σ_X obtained from a single measurement of the anode wire signals in a multi-wire proportional chamber is given by the separation s of the wires: $\sigma_x = s/\sqrt{12}$. The minimum practical s for small chambers is 1 mm. The best resolution is thus about 300 μm.

Significantly better resolution may be obtained either from 'centre of gravity' determination or from the electron drift times in drift chambers.

4.2.5.1 'Centre of Gravity' Method

In this method one uses the signals induced on cathode strips or pads, see Fig. 4.19. The rms width of the induced charge distribution is comparable to the anode-cathode gap d. If one chooses a strip width of (1–2)d, one obtains signals above threshold on typically 3–5 strips. Depending on the signal to noise ratio, a resolution of typically (1–5)% of the strip width is achieved, i.e. about $40 - 100\,\mu\text{m}$. This method is used for the read out of TPCs, as well as for high precision cathode strip chambers, see e.g. [15, 16].

4.2.5.2 Drift Time Measurement

The main contributions to the error of the drift time determination come from electronics noise, electron clustering, δ-rays, diffusion. This assumes that additional effects on the space-time correlation including magnetic field corrections, gain variations, gas contamination and others are kept small by careful construction and calibration. Figures 4.16 and 4.24 (*right*) show typical results. Electronic noise

contributes a constant error. Near the anode wire, the effects of the clustering of the primary charges adds a significant error. At large distances from the anode, the contribution from diffusion grows as square root of distance. Resolutions achieved are typically $50 - 200\,\mu\text{m}$.

A detailed discussion of limits to space resolution is presented in [19] and for the particular case of proportional tubes in [15].

4.2.6 Ageing of Wire Chambers

Deterioration of performance with time has been observed since the early days of gas detectors but has gained importance with the ever increasing radiation loads due to the demand for higher detection rates over long periods. Typical effects of *ageing* are: pulse height decrease, a broadening of the energy resolution and increase in dark current, in the extreme also electrical breakdown or broken wires. An enormous number of studies has been carried out. They are well documented in the proceedings of workshops [54] and several reviews [55].

Upon opening of damaged chambers, deposits have been observed on anode wires and/or on cathodes. On the wire they can take any form from smooth layers to long thin whiskers [56], see Fig. 4.15. On the cathodes, deposits usually consist in spots of thin insulator. Defects of this latter kind can often be correlated with a discharge pattern, which may be interpreted as Malter effect [57]: under irradiation, charges build up on the insulator until the electric field is strong enough to extract electrons from the cathode through the layer into the gas where they initiate new avalanches. The facts that the buildup time decreases with higher ionization rate and that the discharges take some time to decay after irradiation is timed off, give support to this explanation, as does the observation that addition of water vapour is reducing the discharges, probably introducing some conductivity.

Analysis of the layers and whiskers on the anode wires often indicate carbon compounds, more surprisingly also often silicon, sometimes other elements: Cl, O, S.

The aging results are often characterized by a drop in pulse height ΔPH as function of integrated charge deposition in Coulomb per cm wire, although it was found in some cases that the rate of the charge deposition has an influence. Typical values with classical gas mixtures containing hydrocarbons are

$$\begin{aligned} \Delta PH/PH &\sim 0.01 - 0.1\%/0/\text{mC/cm} \quad \text{for small detectors,} \\ \Delta PH/PH &\approx 0.1 - 1\%/\text{mC/cm} \qquad\quad \text{for large detectors.} \end{aligned} \tag{4.63}$$

It is obvious that the control of ageing is one of the major challenges for the LHC experiments, possibly even the major one.

Unfortunately, however, it has not been possible to establish a common fundamental theory, which could predict lifetimes of a new system. On the other hand, the

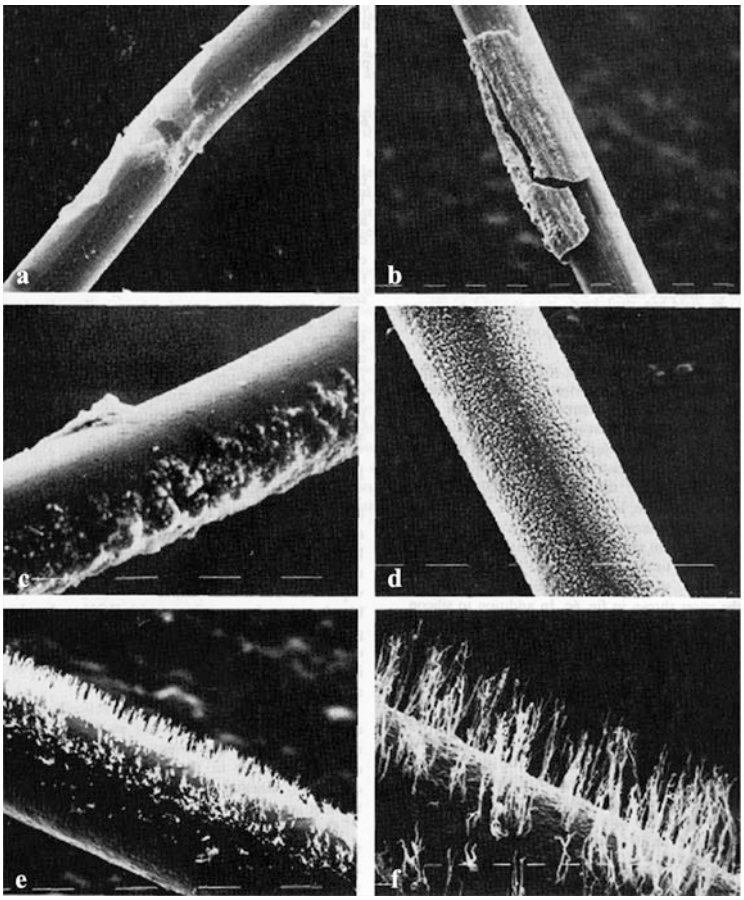


Fig. 4.15 Examples of deposits on 20 μm anode wires after strong irradiation [56]

reasons for ageing in particular circumstances have been elucidated and the studies permit to establish *some general rules* on how for improving the chances for a longer lifetime:

- *Many materials have to be avoided*, in the gas system, in the detector and during the construction: Si compounds, e.g. in bubbler oils, adhesives, vacuum grease or protection foils, PVC tubing, soft plastics in general, certain glues and many more. The workshop proceedings and reviews mentioned present details, also on materials found acceptable.
- For the highest radiation loads, up to a few C/cm, gas components most frequently used in the past, namely *hydrocarbons* like Methane, Ethane, etc., *should be avoided*. Indeed, the LHC experiments make use of them only exceptionally. There remains only a very restricted choice of acceptable gases: mixtures of rare gases and CO_2 and possibly N_2, CF_4 or DME. CF_4 is offering high electron drift

velocities and has proven to be capable under certain conditions to avoid or even to etch away deposits, in particular in the presence of minute Si impurities. But its aggressive radicals may also etch away chamber components, especially glass [15]. In any case the water content has to be carefully controlled to stay below 0.1%, if CF_4 is used, to avoid etching even of gold-plated wires. Also DME, offering low diffusion, has in some cases provided long lifetime. It has, however, shown to attack Kapton and to be very sensitive to traces of halogen pollutants at the ppb level.

- During production, *high cleanliness* has to be observed, e.g. to avoid resistive spots on the cathodes. The sense wire has to be continuously checked during wiring to assure the required quality of its geometrical tolerances and of the gold plating.
- The gas amplification should be kept as low as possible.
- In any case, a final detector module with the final gas system components should be extensively tested under irradiation. As an accelerated test is usually required for practical reasons, to obtain the full integrated charge for some 10 years of operation in a 1 year test, an uncertainty un- fortunately will remain, because a rate dependence of the ageing cannot be excluded.

4.3 Detector Designs and Performance

4.3.1 Single Wire Proportional Tubes

Despite the revolution started with the multiwire proportional chambers (MWPC), single wire tubes are still widely used, mostly as drift tubes. They have circular or quadratic cross-section and offer independence of the cells, important, e.g., in case wire rupture. We present three examples.

ATLAS has chosen for the muon system *large diameter* (*3 cm*) *aluminum tubes* operated with Ar/CO_2 (93/7%) at 3 atm, with the addition of about 300 ppm of water to improve HV stability. A pair of 3 or 4 staggered tube layers, separated by a support frame, form a module. The disadvantages of the gas mixture, a non-linear space-drift time relation and relatively long maximum drift time, had to be accepted in order to obtain a high radiation tolerance. The spatial resolution for a single tube under strong γ- irradiation producing space charge is shown in Fig. 4.16. An average resolution per tube of 80 μm is expected with a maximum background rate of 150 hits/cm^2 s. With a relative positioning of the wires during construction to 20 μm, an adjustment of the tube curvature to the gravitational wire sag and a relative alignment and continuous monitoring of the pair of layers inside a chamber, a combined resolution for the 6–8 1ayers of ~35 μm is aimed at. These chambers provide only one coordinate, the other one being measured in other subdetectors of the experiment.

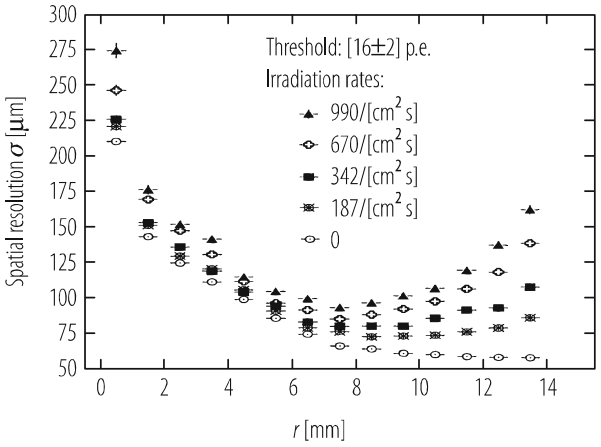


Fig. 4.16 ATLAS MDT drift tubes: space resolution as function of impact radius and background rates. Expected rates are $\leq$150 hits/cm^2 s [15]

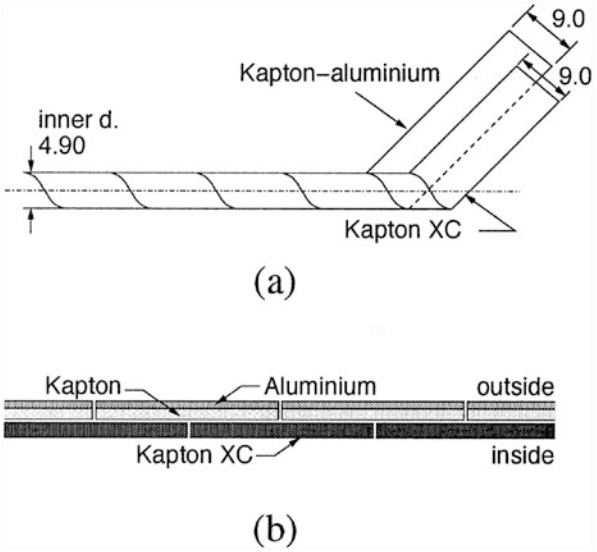


Fig. 4.17 LHCb straw tubes: (**a**) Winding scheme. (**b**) Details of the double foil. Kapton XC is on the inside [17]

A second type of tube design, *straw tubes*, has become very popular since a number of years. Straw tubes offer high rate capability due to small diameters and relatively little material in the particle path. In LHCb, where local rates (near one end of the wire) up to 100 kHz/cm have to be handled, an internal diameter of 4.9 mm was chosen [17], with a construction shown in Fig. 4.17. Two strips of thin foils are wound together with overlap. The inner foil, acting as cathode, is made

of 40 μm carbon doped polyimide (Kapton-XC), the outer is a laminate of 25 μm polyimide, to enhance the gas tightness, and 12.5 μm aluminum, to ensure fast signal transmission and good shielding. The tubes are up to 2.5 m long and have the 25 μm wire supported every 80 cm.

Staggered double layers tubes are glued to light support panels to form modules up to 5 m long. An average spatial resolution of a double layer below 200 μm was measured with Ar/CO_2 (70/30%). In a station, 4 double layers are aligned along $0, +5, -5, 0^\circ$, thus providing a crude second coordinate measurement.

In another design for ATLAS [15], mechanical strengthening of 4 mm straws is achieved with carbon fibers wound around the tubes and straightness by supporting them every 25 cm with alignment planes vertical to the straws. This construction reduces the material along the radial tracks, which is essential for the role of the straws to detect transition radiation originating in fibers stacked in between the tube layers. This role also determines the need for Xe in a mixture of $Xe/CO_2/O_2$ (70/27/3%), the oxygen addition increasing the safety margin against breakdown.

Another quasi-single wire design is that of *plastic streamer tubes* usually called *Iarocci tubes*. Because they are easy to construct in large size and cheap, they have been widely used, especially as readout planes in hadron calorimeters.

A plastic extrusion with an open profile with typically 8 cells of lxl cm^2 and a PVC top plate, is coated on the inside with graphite with a minimum resistivity of 200 kΩ/square. All this is slid into a plastic box, which serves also as gas container. For stability, wires are held by plastic spacers every 50 cm. Using a thick wire of 100 μm, self-quenching streamers are initiated in a gas containing a strong quencher, typically isobutane in addition to Ar. Electrodes of any shape placed on one or both external surfaces pick up the rather strong signals. The dead time is long but only locally, so that particle rates up to 100 Hz/cm^2 can be handled.

4.3.2 Multiwire Proportional Chambers (MWPC)

Already 1 year after the invention of the multiwire proportional counter (MWPC) by Charpak in 1968, a system of small chambers was used in an experiment [58], another year later a large chamber 2 m × 0.5 m had been tested with Ar/Isobutane [59]. A number of developments like bi-dimensional readout were discussed [60]. In 1973 already, a large system of MWPC containing 50,000 sense wires had been constructed for an spectrometer at the ISR, the *Split-Field Magnet* (SFM) [61]. All these chambers had a geometry for the sense wires and gap size similar to the *original design*, shown in Fig. 4.18.

The SFM chambers of 2 × 1 m^2 contained three light support panels forming two amplification gaps of 2 × 8 mm each, one with vertical, the other with horizontal wires of 20 μm diameter. The cathodes on the panels were sprayed with silver paint to provide readout strips 5.5 cm wide and at angles of $\pm 30^\circ$, to resolve ambiguities. Special emphasis was put on high precision with a light construction, including frames of only 5 mm thickness. A total thickness of 1.7% of a radiation length

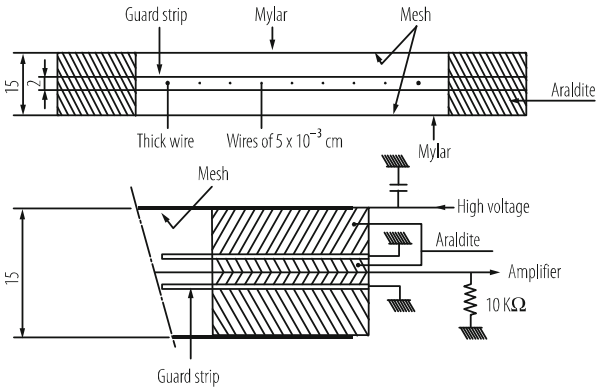


Fig. 4.18 Design of the first multiwire proportional chamber [8]

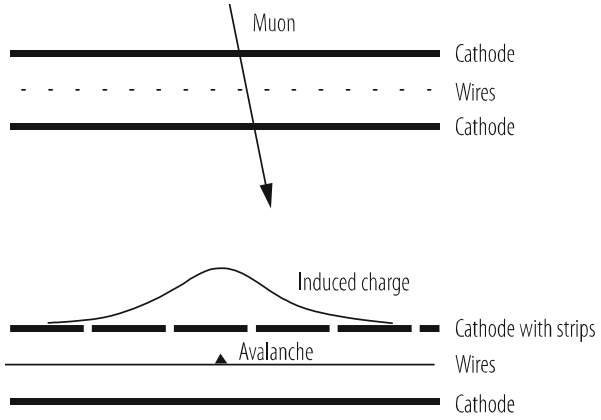


Fig. 4.19 Principle of 'centre of gravity' cathode measurement of cathode strip signals

per chamber was achieved. The stringent quality demands can be inferred from the definition of the *efficiency plateau*: the 'beginning of plateau' was defined as efficiency ε 99.98% and the end by 10 times the 'normal noise', corresponding to cosmics rate With this tight definition, the measured plateau length for a chamber was $50 - 100$ V with the magic gas mixture of Ar/isobutane/freon/methylal (67.6/25/0.4/7%).

In the following decades, drift chambers imposed themselves more and more, but MWPCs remained valid options, especially when speed was more important than high spatial resolution. Thus three of the four LHC experiments employ MWPCs, for triggering and momentum measurements. All of them make use of the signals induced on the cathode strips. In ATLAS and CMS, who call their chambers cathode strip chambers (CSC), a high precision measurement is obtained in the bending coordinate by determining the 'centre of gravity' of the strip signals, see Fig. 4.19. In ATLAS, each third strip is read out (pitch ~5.5 mm), leaving two strips floating,

and a resolution of 60 μm is obtained [15, p. 178]; in CMS, 75 μm resolution is achieved with each strip read out at a minimum pitch of 8.4 mm [16, p. 197]. In LHCb, spatial resolution is secondary to fast timing and high efficiency for a five-fold coincidence trigger. Adjustment to the requirements on spatial resolution, which change strongly with radius, is achieved by forming readout pads of variable size ($0.5 \times 2.5 - 16 \times 20$ cm^2) on the cathodes and by grouping sense wires [17, p. 130].

4.3.3 Drift Chambers

Already in the very first publications, the basic two types of drift chambers were described: (i) with the drift volume, through which the particles pass, separated from the amplification volume [9] and (ii) a geometry, in which the particles pass directly through the volume containing the anode wires alternating with field wires to improve the drift field [10], see Fig. 4.20.

The first design finally evolved into the TPC, the second into a large number of different designs. One can differentiate between planar and cylindrical geometries.

4.3.3.1 Planar Geometries

Most planar geometries are rather similar to each other. To obtain a more homogeneous drift field, additional field shaping electrodes are introduced, see Fig. 4.21. Also shown is a recent example, one element of a layer for the Barrel Muon system of CMS. The space resolution per layer is about 250 μm. One muon station consists of 2×4 layers of such tubes fixed to an aluminum honeycomb plate. The other coordinate is provided by a third set of 4 layers oriented at 90°.

The central detector of UAl used a special arrangement, see Fig. 4.22. In a horizontal magnetic field, at right angle to the beam, a cylinder 6 m long and 2.2 m in diameter is filled with planar subelements. In the central part, vertical anode planes

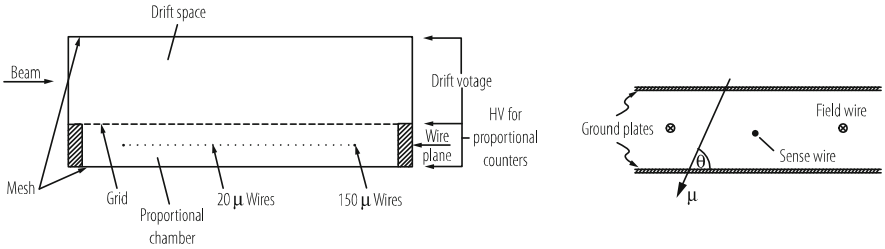


Fig. 4.20 First two drift chamber designs. *Left*: separate drift and amplification gaps [9]. *Right*: Common drift and amplification volume. The additional field wires improve the drift field [10]

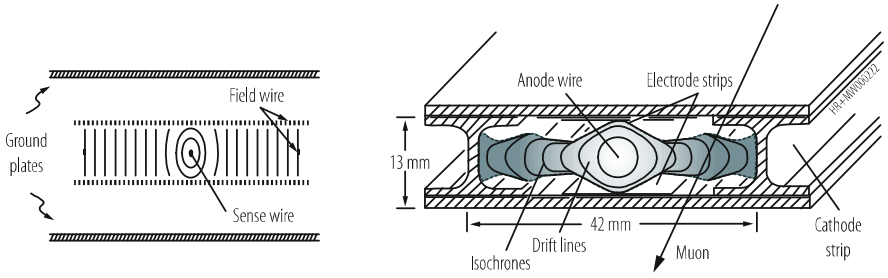


Fig. 4.21 Planar arrangement with field shaping electrodes. *Right*: cross-section of large drift tubes for CMS [16]

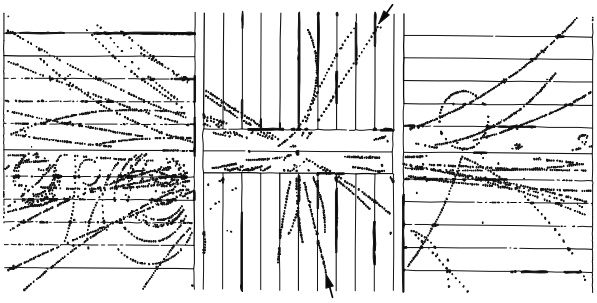


Fig. 4.22 Horizontal view of a reconstructed event in the central detector of UAl. The first Z° decay observed [19]

ultimate with cathode planes, leaving 18 cm drift spaces. These planes are horizontal in the two ends. Charge division is used for the coordinate along the wire. The average point accuracy along the drift direction was 350 μm.

4.3.3.2 Cylindrical Geometries

Many different arrangements have been worked out. Figure 4.23 shows an example of a wire arrangement and electron drift lines following the electric field in the absence of a magnetic field. The change of the electron drift in a magnetic field in a similar cell is indicated to the right.

Figure 4.24 (*left*) shows the conceptual design of the drift chamber for OPAL [53], a wire arrangement called *Jet Chamber*. The left-right ambiguity is solved by staggering the sense wires alliteratively by ±100 μm. The measured space resolution in $r\phi$ for a single wire is presented in Fig. 4.24 (*right*). The figure shows the typical dependence on the distance r from the sense wire: for small r, the primary ion statistics dominates, at large r diffusion. In addition, there is a constant contribution from the noise of the electronics. The coordinate along the wires is

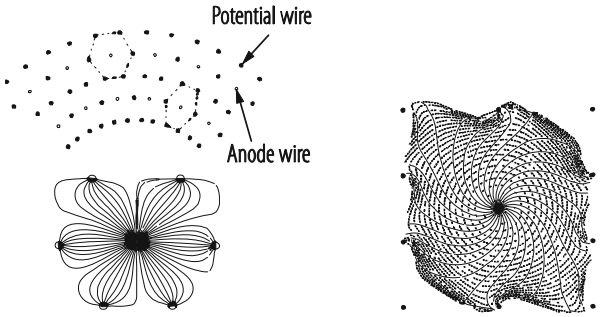


Fig. 4.23 Two multi-layer wire arrangements and electron drift lines without (*left*) and with (*right*) magnetic field

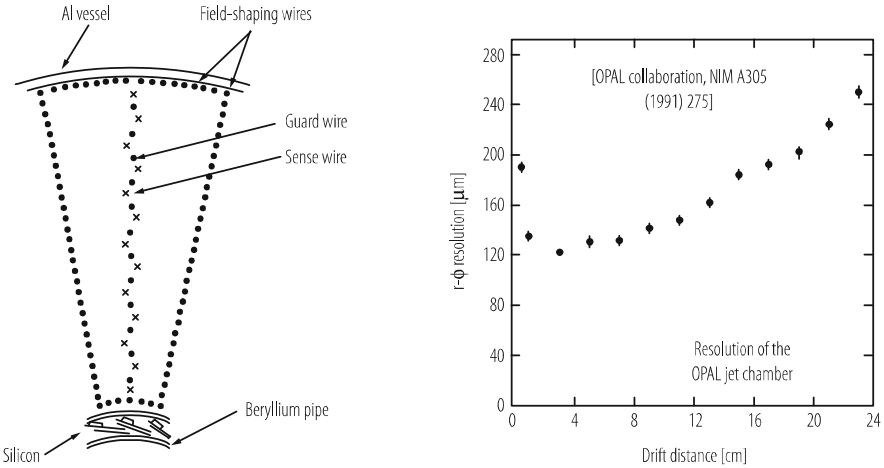


Fig. 4.24 Jet Chamber. *Left*: conceptual design with staggered sense wires. *Right*: Space resolution obtained in OPAL with 4 atm [53]

obtained from *charge division* by using resistive sense wires and read-out on both ends of the wires: a resolution of about 1% of the wire length is reached.

In other designs the *second coordinate* is obtained from orienting successive layers in stereo angles. Sometimes relative timing with read-out of both ends of the wire is used, providing again a resolution of about 1% of the wire length.

4.3.3.3 Time Projection Chambers (TPC)

The TPC concept proposed by Nygren [62] in 1974 for the PEP4 experiment [63] offered powerful pattern recognition with *many unambiguous 3-D points* along a track and particle identification by combining dE/dx information from many samples with momentum measurement. Originally proposed to resolve jets at a low

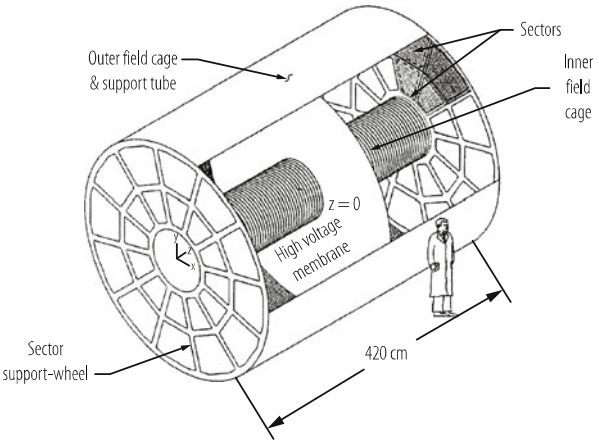


Fig. 4.25 Conceptual design of the STAR TPC operating at RIC [64]

energy e^+e^- collider, the TPC design has proven years later to be the most powerful tracker to study central heavy ion collisions with up to several thousand particles in an event, at more than 100 events per second.

The basic design elements have hardly changed over the years. Cylindrical field cages provide a homogeneous electric field between the central electrode and the planar wire chambers at both ends; see Fig. 4.25 for the conceptual design of the latest TPC in operation, the STAR TPC at RHIC [64]. The typical gas mixture is Ar/CH_4, which offers high drift velocity at low electric field and low electron attachment. The electrons from the track ionization drift to one of the two endcaps. They traverse a gating grid and a cathode grid before being amplified on 20 μm anode wires, separated with field wires. The avalanche position along the wires is obtained from measuring the centre of gravity of pulse heights from pads of the segmented cathode beneath. Figure 4.26 shows the electric field lines for a closed and an open gating grid. *Gating* is essential for the TPCs with their long drift length, to reduce space charge build-up. The gate is only opened on a trigger.

All TPCs except PEP4 and TOPAZ operated at latm and profit from a strong reduction of lateral diffusion due to the factor $\omega\tau$~5 in the strong magnetic field B oriented parallel to the electric field E. Higher pressure is rather neutral: $\omega\tau$ decreases, but more primary electrons reduce relative fluctuations and thus **ExB** and track angle effects. Typical point resolutions in $r\Phi$ range from 150 to 200 μm at the e^+e^- colliders [65]. Figure 4.27 shows a reconstructed Pb–Pb interaction observed in STAR.

All TPCs except STAR and ALICE use the signals from the anode wires for dE/dx information. In STAR and ALICE, all information is taken from the pads, some 560,000 in ALICE [14]. Pressure improves dE/dx and the PEP4 TPC operating at 8.5 atm produced the best dE/dx resolution despite a smaller radius [65].

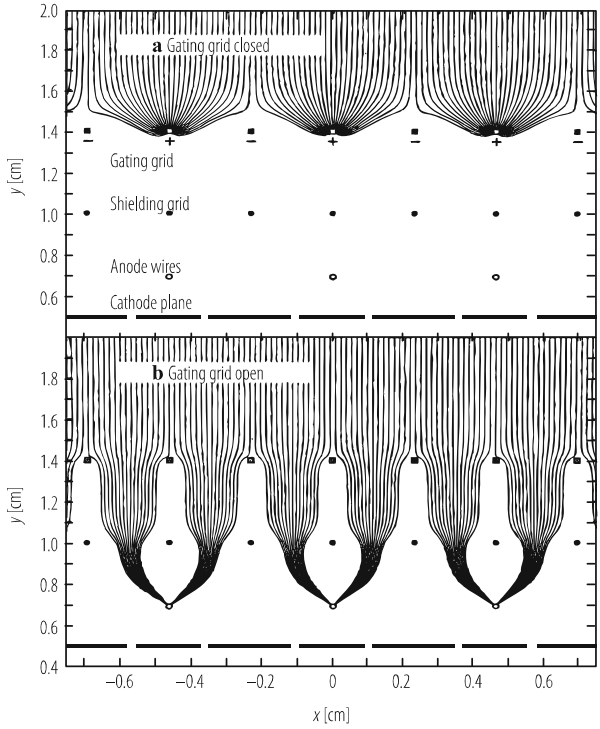


Fig. 4.26 TPC wire chamber: electric field lines for a closed (**a**) and open x [cm] gating grid **(b)**

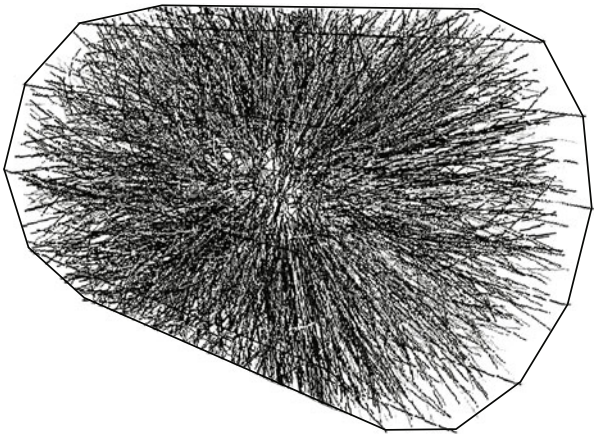


Fig. 4.27 A reconstructed Pb–Pb interaction observed in the STAR TPC [64]

4.3.4 Parallel Plate Geometries, Resistive Plate Chambers (RPCs)

Parallel plate devices offer fast response, as there is no drift delay and the avalanche amplification starts immediately.

Keuffel's spark counter [6] featured two metal electrodes at millimeter distance in a gaseous atmosphere, where the primary electrons deposited in the gap provoke a fast discharge and therefore a detectable signal. By the end of the 1960s the spark counters had arrived at time resolutions around 100 ps, the rates however were limited to 1 kHz and small areas of about 30 cm^2, since after each discharge the entire counter was insensitive during the recharge time of typically a few hundred microseconds. Parallel Plate Chambers (PPCs) use the same geometry but operate below the discharge voltage. The avalanche therefore induces a signal but does not create a discharge, allowing a rate capability of 100 kHz for a 80 cm^2 detector [6]. Still, the fact that the detector mechanics and especially the detector boundaries have very carefully controlled to ensure stability, limits this detector to a rather small surface.

The Pestov spark counter [66] uses the same parallel plate geometry, with one electrode made from resistive material having a volume resistivity of $\rho = 10^9 - 10^{10}$ Ωcm. The charge deposited locally on this resistive layer takes a time of $\tau \approx \rho\varepsilon$ to be removed, where ε is the permittivity of the resistive plate. This time is very long compared to the timescale of the avalanche process, the electric field is therefore reduced at the location of the avalanche, avoiding a discharge of the entire counter. This allows stable operation of the detector at very high fields and particle rates. A counter with a size of 600 cm^2 and a gas gap of 1 mm, operated at atmospheric pressure achieved a time resolution of <0.5 ns and efficiency of 98%. By decreasing the size of the gas gap to 0.1 mm and operating the detector at 12 bar pressure, a time resolution of 27 ps was achieved with this detector [67].

Resistive Plate Chambers (RPCs) [68] are building on this same principle and they are widely used as trigger detectors and for time-of-flight measurements, as they allow relatively cheap large area construction. Large detector systems of several hundred m^2 surface have been built with Bakelite plates ($\rho = 10^{10} - 10^{12}$ Ωcm) or window glass ($\rho = 10^{12} - 10^{13}\,\Omega$ cm). Tetrafluorethane ($C_2F_4H_2$) is nowadays widely used as the main component of the RPC gas mixture due to the large number of primary ionization clusters (8–10/mm) leading to large detection efficiency and due to it's electronegativity that reduces the probability for the formation of streamers. Small additions of SF_6 are further reducing this streamer probability.

The time resolution for RPCs is given by $\sigma_t \approx 1.28/\alpha v$, where α is the effective Townsend coefficient of the gas mixture and v is the drift-velocity of the electrons. RPCs with a single gas gap of 2 mm at a field of 50 kV/cm ($\alpha \approx 10$/mm, $v \approx 130\,\mu$m/ns) provide a time resolution of $\approx$1 ns and efficiency close to 100%. Figure 4.28a shows the geometry as used for the muon system of the ATLAS experiment. In addition to collider experiments, similar geometries are used as

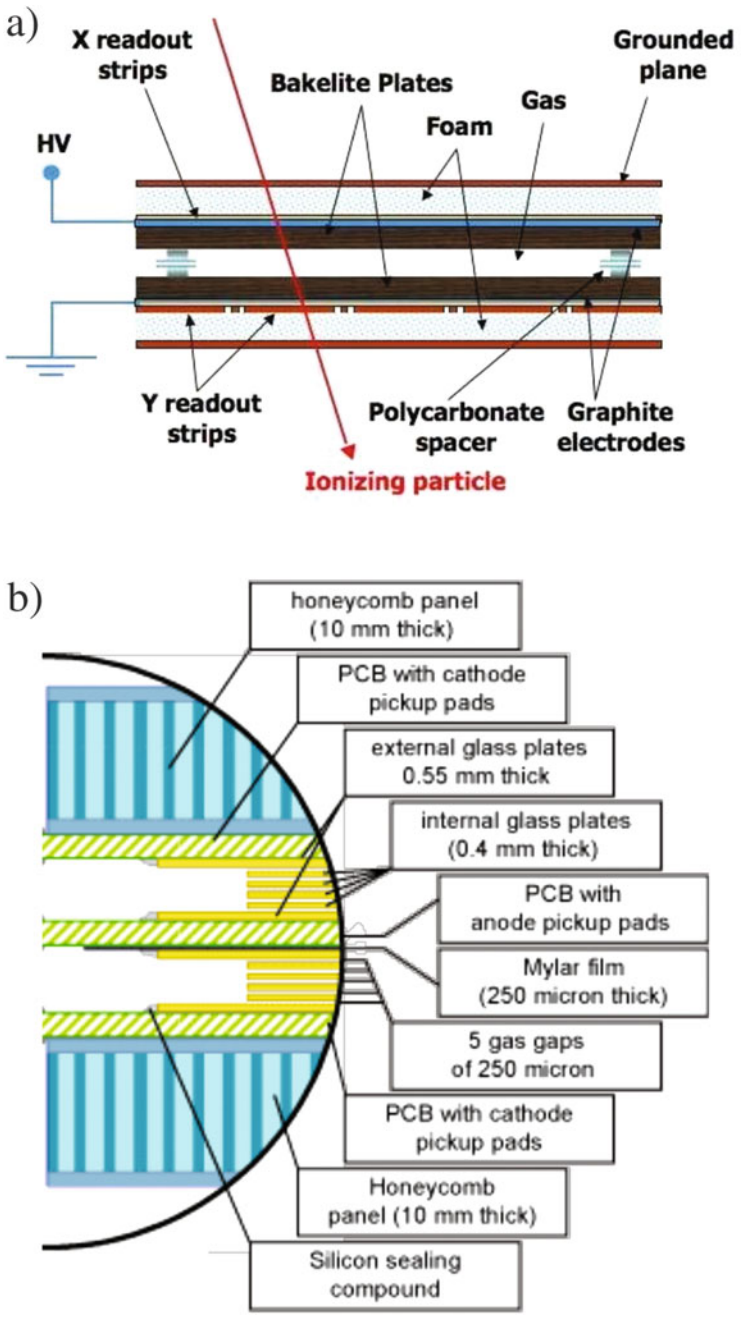


Fig. 4.28 (**a**) Single gap RPC as used by the ATLAS experiment for the muon trigger system. (**b**) Multigap RPC as used by the ALICE experiment for the time of flight system

trigger or veto detectors in neutrino experiments like OPERA [69] and Daya Bay [70] and as large area cosmic ray detectors like ARGO [71].

Using a small gas gap of 0.25–0.3 mm with a field around 100 kV/cm ($\alpha \approx 113$/mm, $v \approx 210\,\mu$m/ns) results in a time resolution of $\approx$50 ps, making the detector well suited for time-of-flight measurements. The reduced efficiency due to the narrow gas gap is overcome by using a multi-gap structure [83]. Figure 4.28b shows the geometry as used for the time of flight system of the ALICE experiment. The avalanche process in RPCs is significantly affected by spacecharge effects. After the initial exponential increase of the electron number, the ions produced in the avalanche are significantly reducing the electric field and therefore resulting in strong slow down of the avalanche growth. This results in moderate signal charges in the pC range even for very large Townsend coefficients [72].

The rate capability of RPCs is defined by the thickness d of the resistive plates and their volume resistivity ρ. The current I produced per unit area inside the gas gap is flowing through these plates, which results in an effective voltage drop of $\Delta V = I\rho d$ across a single plate. The rate limit of the RPC is reached at the point where the effective voltage across the gas gap moves outside the efficiency plateau. For the values and geometries quoted above, this limit is in the range of 10–1000 Hz/cm^2.

4.3.5 Micropattern Devices

The constantly increasing particle rates and track densities in modern day experiments exceed the capabilities of standard gaseous detectors. Semiconductor technology dominates this regime. On the other hand, numerous novel designs of gaseous detectors have been studied. Two have emerged and attract much attention, the so-called GEM and Micromegas devices. Offering small ExB track distortions and low ion feedback, they are also being used for the readout of TPCs.

4.3.6 Gas Electron Multiplier (GEM)

In a thin metal-coated polymer foil, holes are chemically etched at high special density [73], see Fig. 4.29. A voltage applied to the metal layers produces gas amplification in the holes. Typical parameters are: Foil thickness = 50 μm, inner hole diameter = 70 μm, hole pitch = 140 μm, voltage = 400 V. To achieve a practical gas gain of the order of $10^4 - 10^5$ with an acceptable low discharge probability, usually three GEMs are put in series. In COMPASS [74], a system of 20 triple-GEMs with an active area of 31 $\times$ 31 cm^2 each was operated in a very high intensity muon beam. With 2-D readout via superposed orthogonal strips, a space resolution of 70 μm was achieved at rates up to 2.5 MHz/cm^2. The efficiency was 99% with 50 ns pulse shaping at an effective gain of 8000.

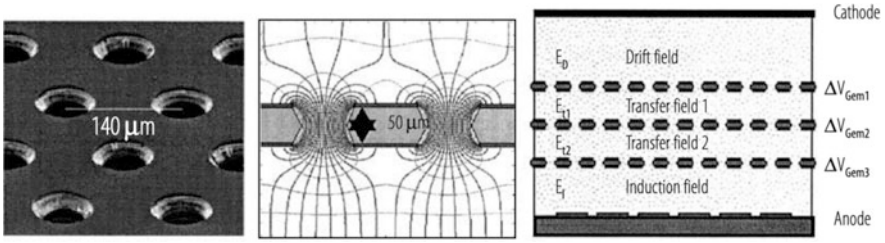


Fig. 4.29 GEM. Left: hole structure. Centre: electric field lines. Right: Conceptual design of tripleGEM [73]

In LHCb [75], with a much shorter peaking time of 10 ns, an efficiency of $\geq$96% was reached for two triple-GEMs in OR and a gain of 6000 with $Ar/CO_2/CF_4$ (45/15/40). The time resolution with this chamber was $\leq$3 ns with Ar/CO_2 (70/30).

Charge build-up is observed on the insulating holes in the GEM foils, which varies with the particle flux and is accompanied by some change in gain, well described by simulation [76].

Several large scale GEM systems are under construction. The wire chambers of the ALICE TPC are replaced by an a quadruple-GEM arrangement that is optimized for low ion backflow. This allows the TPC to operate in continuous mode without the need for gating. The top and bottom GEMs have a hole separation of 140 μm as described above, while the two middle GEMs have twice this distance. This results in an ion feedback below 1% and energy resolution of <12% for 5.9 keV photons at an effective gain of $\approx$2000 [77].

The muon system of the CMS experiment is being equipped with standard triple-GEM detectors over a surface of $\approx$220 m^2 providing spatial resolution of $200 - 400\,\mu$m and a time resolution of 8 ns [78].

4.3.7 Micromegas

In a Micromegas detector [79], the ionization produced in the drift gap is channeled through an extremely fine mesh into the amplification gap, terminated by an anode plane segmented into readout strips or pads as seen in Fig. 4.30. The 'micromesh' is woven from $\approx 15\,\mu$m wires leaving holes of about 50 μm^2. The amplification gap is only 50–150 μm thick and behaves on average like a parallel counter. The mesh is supported by pillars every $\approx$2.5 mm. Due to the high amplification field in comparison to the drift field, the electrons are moving to the anode only inside a very thin funnel, see Fig. 4.30.

In the COMPASS experiment, 12 chambers 40×40 cm^2 have operated in fluxes up to 25 MHz/mm^2 and obtained resolutions of 70–90 μm and 9 ns. The near detector of the T2K experiment [75] uses a TPC with 9 m^2 of MICROMEGAs detectors with readout pads of 10×7 mm^2. The muon system of the ATLAS detector is

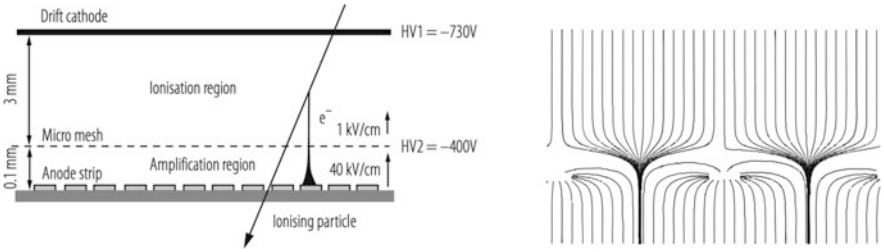


Fig. 4.30 Micromegas. Left: conceptual design. Right: Electrical field lines

implementing two 'wheels' of 8 m diameter with 4 layers of MICROMEGAs [80]. The readout readout strips of 300 μm width achieve a position resolution around 100 μm. In order to increase the stability against discharges for these very large surfaces, resistive strips are placed on top of the readout strips at a distance of 64 μm. The resistance value of 10–20 MΩ/cm ensures that the rate capability is sufficient for the application.

4.4 Outlook

The availability of large area silicon sensors has allowed most of the recent detector setups to use silicon trackers for vertexing and momentum spectroscopy in the detector volume upstream of the calorimeter systems. Muon systems do however have surfaces of up to several thousands of m^2 with particle rates and resolution requirements that make the application of gas detectors still the most viable solution. The TPC is still a very appealing detector for setups where very low material budget as well as PID capabilities are important requirements. Experiments such as NEXT [81] for the search of neutrinoless double beta decay are building on the unique features of gas detectors like low density of the detection medium and the related possibility for tracking of very low energy particles. Gas detector will therefore continue to be essential elements of particle physics instrumentation.

The last two sections on Resistive Plate Chambers and Micropattern Devices were updated in this new edition, while the remainder of this chapter is in its original form by H.J. Hilke.

References

1. E. Rutherford, H. Geiger, Proc. Roy. Soc. London A 81 (1908) 141.
2. H. Geiger, Verh. D. Phys. Ges. 15 (1913) 534.
3. H. Greinacher, Z. Phys. 23 (1924) 261.
4. H. Geiger, W. Mueller, Phys. Z. 29 (1928) 839.

5. A. Trost, Z. Phys. 105 (1937) 399.
6. (a) J.W. Keuffel, Rev. Sci. Instrum. 20 (1949) 202; (b) J. Christiansen, Z. Angew. Phys. 4 (1952) 326.
7. E. Fuenfer, H. Neuert, Zählrohre und Szintillationszähler, G. Braun, Karlsruhe (1959).
8. G. Charpak et al., *The Use of Multiwire Proportional Counters to select and localize Charged Particles*, Nucl. Instrum. Meth. 62 (1968) 262.
9. T. Bressani, G. Charpak, D. Rahm, C. Zupancic, *Track Localization by Means of a Drift Chamber*, Proc. Int. Seminar Filmless Spark and Streamer Chambers, Dubna, USSR, 1969, p. 275.
10. A.H. Walenta, J. Heintze and B. Schürlein, *The Multiwire Drift Chamber-A Novel Type of Proportional Wire Chamber*, Nucl. Instrum. Meth. 92 (1971) 373.
11. R. Veenhof, *Garfield-simulation of gaseous detectors*, http://consult.cem.ch/writeup/garfield/
12. I. Smimov, *Heed*: *Interactions ofparticles with gases*, http://consult.cem.ch/writeup/heed/
13. S. Biagi, *Magboltz*: *Transport of electrons in gas mixtures*, http://consult.cem.ch/writeup/magboltz/
14. The ALICE experiment at the CERN LHC, J. Instrum. 3 (2008) S08001.
15. The ATLAS Experiment at the CERN LHC, J. Instrum. 3 (2008) S08002.
16. The CMS experiment at the CERN LHC, J. Instrum. 3 (2008) S08003.
17. The LHCb Detector at the LHC, J. Instmm. 3 (2008) S08004.
18. Special volumes of Nucl. Instrum. Methods: the last three Proceedings are found in Nucl. Instrum. Meth. A 581 (2007) 1; 535 (2004) 1; 478 (2002) 1.
19. W. Blum, W. Riegler, L. Rolandi, Particle Detection with Drift Chambers, 2nd ed., Springer-Verlag (2008).
20. H. Bichsel, (a) Nucl. Instrum. Meth. A 562 (2006) 154; (b) Phys. Lett. B667 (2008) 267.
21. Review Particle Physics, Phys. Lett. B667(1–5) (2008) 292.
22. F. Lapique, F. Piuz, Nucl. Instrum. Methods 175 (1980) 297
23. H. Fischle, J. Heintze, B. Schmidt, Nucl. Instrum. Methods 301 (1991) 202
24. H. Walenta, Nucl. Instrum. Methods 161 (1979) 45.
25. I. Lehraus et al., Nucl. Instrum. Methods 153 (1978) 347
26. J. Va'vra et al. SLAC-PUB-5728, (1992).
27. S.F. Biagi, Nucl. Instrum. Meth. A 283 (1989) 716; ref. to J.C. Armitage, Nucl. Instrum. Meth. A 271 (1988) 588.
28. B. Schmidt, Doctoral thesis, Univ. Heidelberg, (1986).
29. G. Schultz, Doctoral thesis, Univ. Louis-Pasteur, Strasbourg, (1979).
30. Review Particle Physics, Physics Letters. B667(1–5) (2008) 1.
31. Compilation by A. Jeavons et al., Nucl. Instrum. Meth. 176 (1980) 89–97.
32. J.H. Hombeck, Phys. Rev. 84 (1951) 615.
33. (a) RD-32, CERN/DRDC 94-10 (1994); (b) A. Ishikawa, http://www-hep.phys.saga-u.ac.jp/ILC-TPC/gas/ps/
34. E.B. Wagner, F.J. Davies, F.J. Hurst, J. Chem. Phys. 47 (1967) 3138.
35. J.H. Parker, J.J. Lowke, Phys. Rev. 181 (1969) 290 and 302
36. PEP-4 Proposal, SLAC-Pub-5012 (1976).
37. S.R. Amendolia et al., Nucl. Instrum. Meth. 244 (1986) 516.
38. H.S.W. Massey, E.H.S. Burhop, H.B. Gilbody, Electronic and ionic impact Phenomena, Clarendon, Oxford (1969).
39. D.L. McCorkle, L.G. Christophorou, S.R. Hunter, in: Proc. 2nd Int. Swarm Seminar, Oak Ridge, USA, Pergamon, New York (1981) p. 21.
40. F. Bloch, N.E. Bradbury, Phys. Rev. 48 (1935) 689.
41. M. Atac, A.V. Tollestrup, D. Potter, Nucl. Instrum. Meth. 200 (1982) 345.
42. Collected by A. von Engel, Handbuch der Physik 21 (1956), Springer, Berlin.
43. M.E. Rose, S.A. Korff, Phys. Rev. 59 (1941) 850.
44. W. Diethom, USAEC Report NY 6628 (1956).
45. G.A. Erskine, Nucl. Instrum. Meth. 105 (1972) 565.
46. C. Brand et al. Nucl. Instrum. Meth. A 237 (1985) 501.

47. W. Legler, Z. Naturforschung A 16 (1961) 253.
48. H. Schlumbohm, Z. Phys. 151 (1958) 563.
49. G.D. Alkahazov, Nucl. Instrum. Meth. 89 (1970) 155.
50. S.C. Curran, A.L. Cockroft, J. Angus, Philos. Mag. 40(1949) 929.
51. S. Ramo, Proc. I.R.E. 27(1939) 584.
52. W. Shockley, J. Appl. Phys. 9 (1938) 635.
53. OPAL Collaboration, *The OPAL Detector at LEP*, Nucl. Instrum. Meth. A 305 (1991) 275.
54. (a) Proc. Workshop Radiation Damage to Wire Chambers, Berkeley (org. J. Kadyk), LBL 21170 (1986); (b) Proc. Int. Workshop Aging Phenomena in Gaseous Detectors, Hamburg, 2001, Nucl. Instrum. Meth. A 515 (2003); (c) NASA, a very large database on outgassing of materials can be reached via the NASA home- page.
55. (a) J. Va'vra, Nucl. Instrum. Meth. A 515 (2003) 263, and N. Tesch, IEEE Trans. Nucl. Sci. 49 (2002) 1609; (b) M. Capeans, Nucl. Instrum. Meth. A 515 (2003) 73; (c) R. Bouclier et al., Nucl. Instrum. Meth. A 350 (1994) 464.
56. J. Adam et al., Nucl. Instrum. Meth. 217 (1983) 291.
57. L. Malter, Phys. Rev. 50 (1936) 48.
58. C. Bemporad, Nucl. Instrum. Meth. 80 (1969) 205.
59. P. Schilly et al., Nucl. Instrum. Meth. 91 (1970) 221.
60. G. Charpak, D. Rahm, H. Steiner, Nucl. Instrum. Meth. 80 (1970) 13.
61. R. Bouclier et al., Nucl. Instrum. Meth. 115 (1973) 235
62. D.R. Nygren, Proposal to investigate a novel concept in particle detection, LBL intemal report, Berkeley, February 1974.
63. Proposal for a PEP Facility based on the Time Projection Chamber, PEP 4, Dec. 1976.
64. M. Anderson et al., *The STAR Time Projection Chamber*, Nucl. Instrum. Meth. A 499 (2003) 659.
65. (a) A. Shirahashi et al., *TOPAZ Time Projection Chamber*, IEEE Trans. NS-55 (1988) 414; (b) W.B. Atwood et al., *ALEPH Time Projection Chamber*, Nucl. Instrum. Meth. A 306 (1991) 446; (c) P. Abreu et al., *DELPHI Detector*, Nucl. Instrum. Meth. A 178 (1996) 57; (d) J. Alme et al., *The ALICE TPC*, Nucl. Instrum. Meth. A 622 (2010) 316-367; (e) a review on TPCs: H.J. Hilke, Rep. Prog. Phys. 73 (2010) 116201.
66. V.V. Parkhomchuk, Y.N. Pestov and N.V. Petrovykh, A spark counter with large area, Nucl. Instr. and Meth. A 93 (1971) 269–270.
67. Yu.N. Pestov, Status and future developments of spark counters with localized discharge, Nucl. Instr. and Meth. A 196 (1982) 45–47.
68. R. Santonico, R. Cardarelli, Development of Resistive Plate Counters, Nucl. Instr. and Meth. A 187 (1981) 377.
69. A. Bertolin et al., The RPC system of the OPERA experiment, Nucl. Instr. and Meth. A 602 (2009) 631–634.
70. J. Cao and Kam-Biu Luk, An overview of the Daya Nay reactor neutrino experiment, Nuclear Physics B 908 (2016) 62–73.
71. G. Aielli et al., Layout and performance of the RPCs used in the Argo-YBJ experiment, Nucl. Instr. and Meth. A 562 (2006) 92–96.
72. W. Riegler et al., Detector physics and simulation of Resistive Plate Chambers, NIMA 500 (2003) 144–162.
73. F. Sauli, Nucl. Instrum. Meth. A 386 (1997) 531.
74. B. Ketzer et al., Nucl. Instrum. Meth. A 535 (2004) 314.
75. N. Abgrall et al., Time projection chambers for the T2K near detectors, NIMA 637 (2011) 25–46.
76. V. Tikhonov, R. Veenhof, Nucl. Instrum. Meth. A 478 (2002) 452.
77. M.M. Aggarwal et al. NIM A 903 (2018) 215.
78. D. Abbaneo et al., Upgrade of the CMS muon system with tripple-GEM detectors, JINST 9 C10036.
79. Y. Giomataris, Nucl. Instrum. Meth. A 419 (1998) 239.

80. F. Kuger et al., Performance studies of the resistive Micromegas detectors for the upgrade of the ATLAS Muon spectrometer, NIMA 845 (2017) 248–252.
81. J.J. Gomez-Cadenas, The next experiment, NPPP 273-275 (2016) 1732–1739.
82. IAEA-TECDOC-799 (1995) 560, Atomic and Molecular Data for Radiotherapy.
83. A. Akindinov et al., The multigap resistive plate chamber as a time-of-flight detector, Nucl. Instr. and Meth. A 456 (2000) 16.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 5
Solid State Detectors

G. Lutz and R. Klanner

5.1 Introduction

Semiconductor detectors, and in particular silicon detectors, are very well suited for detection and measurement of light and of ionizing radiation caused by interaction with charged particles and (X-ray) photons. Precise position, time and energy measurement can be combined when use is made of the excellent intrinsic material properties in well thought out detector concepts.

Development and large scale use of silicon detectors has been initiated by particle physics. The discovery of the rare and short lived charmed particles lead to the desire to use their decay topology as signature for identification and separation from non-charm background. Detectors were required that combined very good position measurement (in the range of several μm) with high rate capability (few hundred kHz), a task not achievable with available detectors at that time.

Semiconductor detectors, in particular silicon and germanium detectors were used for quite some time, but not too frequently, in Nuclear Physics for the

In the updated version a number of detector developments which took place after the publication of the original version have been taken into account. These are in particular new sections on Radiation Damage, 3-D Detectors, MAPS (Monolithic Active Pixel Sensors), SiPMs (Silicon Photomultipliers) and Ultrafast Tracking Detectors (LGAD = Low Gain Avalanche Detectors). In addition, the section Summary and Outlook has been updated.

The author G. Lutz is deceased at the time of publication.

G. Lutz
PNSensor GmbH and MPI-Halbleiterlabor, Munich, Germany

R. Klanner (✉)
Department of Physics, University of Hamburg, Hamburg, Germany
e-mail: robert.klanner@desy.de

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_5

purpose of measuring particle and X-ray photon energies, not however for position measurement. This task was left mostly to gas detectors and to scintillation hodoscopes, both of them not able to provide the required position measurement resolution.

It was realized rather soon that semiconductors offer in principle the required capabilities and silicon strip detectors were developed and used for the detection and investigation of charmed particles. This development rapidly increased in speed and scope so that today it is rare to find particle physics experiments that do not rely heavily on silicon strip detectors for particle tracking and identification. Strip detectors have also entered many other fields of science. Important features of this development were the introduction of more sophisticated detector concepts and the development of multi-channel low noise-low power integrated readout electronics adapted to the requirements of strip detectors.

A further challenge in particle tracking poses the ambiguities occurring in case of high particle densities. This problem is alleviated considerably when replacing the strip geometry by pixels. Hybrid pixel detectors became possible with the enormous progress in miniaturization of electronics. Each pixel has its own readout channel. Detector and electronics with matched geometry are connected face to face by bump bonding. Recently Monolithic Active Pixel Sensors (MAPS), pixel detectors in which sensor and readout electronics are integrated on the same silicon chip, are reaching maturity.

Although in the initial phase of this rapid development position measurement was in the focus of interest, energy resolution with high readout speed came back to its right, sometimes in combination with position resolution. This development opened the door of semiconductor detectors in X-ray astronomy, synchrotron radiation experiments and in many other fields.

A major step on this way was the invention by E. Gatti and P. Rehak of the semiconductor drift chamber [1]. This concept also became the basis for further new concepts as are the pnCCD [2], the silicon drift diode [3] and the DEPFET [3] that forms the basis for several types of pixel detectors with rather unique properties.

In the last decade, a major progress in the field of silicon photo-detectors took place: Multi-pixel avalanche photo diodes operating in the Geiger mode, frequently called silicon photo-multipliers, SiPM, have been developed and found many applications in research, medicine and industry.

In the following, detection principles and properties of the various detector types will be described and some applications will be sketched. Emphasis is on detector physics and concepts while it is impossible to cover all important activities in the field. In addition, a short summary of radiation damage, which presents a major challenge for the use of silicon detectors in the harsh radiation environment at colliders, like at the CERN Large Hadron Collider, LHC, will be presented.

5.2 Basic Detection Process of Single Photons in Semiconductors

The simplest detector is a reverse-biased planar diode (Fig. 5.1). Photons interacting in silicon will, dependent on their energy, produce one or more electron-hole pairs close to their points of interaction. Charged particles will generate pairs along their path within the semiconductor. An average energy of 3.6 eV is needed for creation of a pair in silicon with a band gap of 1.12 eV at room temperature. This should be compared with the ionization energy of gases which is more than an order of magnitude higher. Electrons and holes will be separated by the electric field within the space charge region and collected at the electrodes on opposite sides of the diode.

The small band gap and the corresponding large signal charge generated in the photon absorption process is the principal cause for the excellent properties of semiconductor radiation detectors manifesting themselves in particular in very good spectroscopic resolution down to low energies. Further reasons are the high density and corresponding low range of delta-electrons which makes very precise position measurement possible. High charge carrier mobilities combined with small detector volume leads to short charge collection time and makes the use of detectors in high rate environment possible. The excellent mechanical rigidity makes the use of gas containment foils superfluous and allows operation in the vacuum. Therefore very thin entrance windows can be constructed and high quantum efficiency can be reached down to low photon energies. Position dependent doping of semiconductors allows construction of detectors with sophisticated electric field configurations and intrinsically new properties.

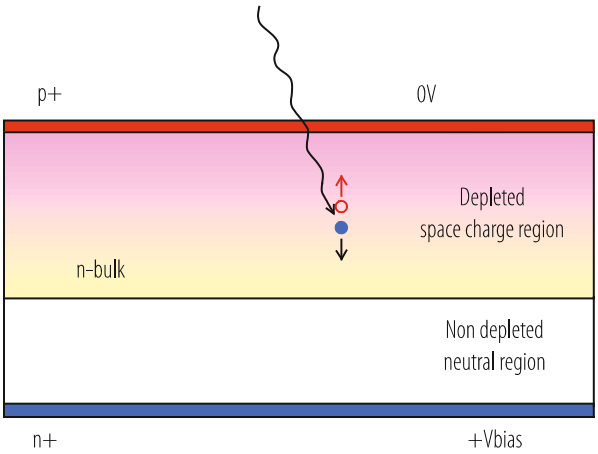


Fig. 5.1 Schematic structure of a reverse-biased semiconductor diode used as photon detector. The region heavily doped with acceptors is denoted p^+, and n-bulk and n^+ the regions lightly and heavily doped with donors, respectively

Reaching all these good detector properties requires a readout electronics which is well matched to the detectors. Here we notice a point specific to silicon which is also the basic material of most of present day electronics. For that reason it is natural to integrate the sensitive front-end part of electronics into the detector. This is the case, for example, in CCDs and drift diodes [3] with very high spectroscopic resolution. A further device (DEPFET) [3] combines the function of detector and amplifier in the basic structure. In MAPS (Monolithic Active Pixel Sensors) sophisticated readout electronics is directly integrated on the silicon chip of the sensor.

Dependent on the field of application different aspects of semiconductors are in the focus of interest. In particle physics tracking requires high position resolution and often high speed capabilities while energy resolution is of less importance. Recently at the CERN LHC also a timing accuracy of a few tens of picoseconds in combination with precision tracking became a requirement. In X-ray spectroscopy and imaging, as well as in X-ray astronomy both energy and position resolution are of importance. For light detection, high photon-detection efficiency and resolving single photons are typically more important than position accuracy.

5.3 Basics of Semiconductor Physics

After these introductory remarks on semiconductor detectors we will look into the underlying mechanisms in a little more detail.

Most commonly used semiconductors are single crystals with diamond (Si and Ge) or zinc blende (GaAs and other compound semiconductors) lattice. Each atom in the crystal shares their outermost (valence) electrons with the four closest neighbours. At very low temperature all electrons are bound to their respective locations and the material is an insulator. At elevated temperature thermal vibrations will sometimes break a bond and both the freed electron and the hole (the empty place left behind to be filled by a neighbouring electron) are available for electrical conduction. The density of free electrons/holes is called intrinsic carrier density n_i. For silicon its value at room temperature is about 10^{10} cm^{-3}, resulting in an intrinsic resistivity of about 350 kΩ·cm.

Creation of electron–hole pairs can also be accomplished by electromagnetic radiation or by the passage of charged particles knocking out of their covalent bond some of the valence electrons. This is the mechanism used in the detection process. These free charge carriers will then be moved by an applied electrical field (drift) and redistribute due to concentration variations (diffusion) until finally reaching an external electrode connected to the readout electronics.

So far we have only dealt with intrinsic semiconductors, perfect crystals without foreign atoms. One may, however, replace a small fraction of atoms with some having either one more, called donors (e.g. P in Si) or one less, called acceptors (e.g. B in Si) valence electron. The additional electron or the missing electron (hole) is only weakly bound, resulting in states in the silicon band gap located about 40 meV

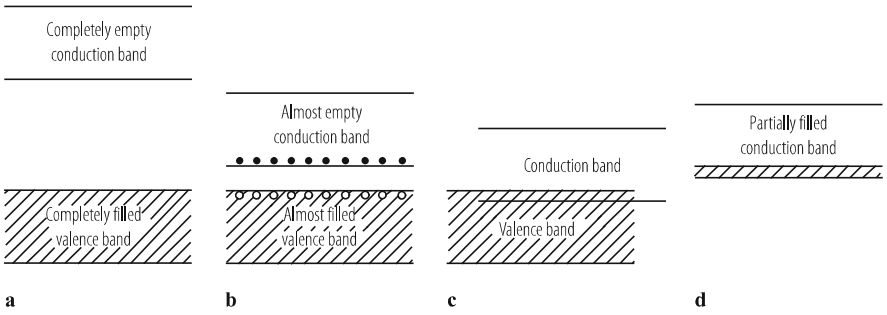


Fig. 5.2 Energy band structure of insulators (**a**), semiconductors (**b**), and conductors (**c**, **d**)

from the conduction or valence band, respectively. silicon doped with donor atoms is called *n*-type, and *p*-type for acceptors. These states are already ionized well below room temperature, and the electrons or holes can move freely in the silicon lattice, resulting in a decrease of the resistivity. For silicon detectors crystals with a typical doping density of 10^{12} cm^{-3} are used, which results in a similar density of free charge carriers and a significantly reduced resistivity of a few kΩ·cm. Applying an external electric field the free charge carriers can be removed and a space charge region due to the surplus charge of the doping atoms is created.

The discussion so far has used the simple bond picture. A more sophisticated treatment that allows also quantitative calculations requires the quantum mechanical band model. While single atoms possess discrete energy levels, in crystals these are transformed into energy bands.

Figure 5.2 shows the (almost) fully occupied valence band and the lowest laying (almost) empty conduction band for insulators, semiconductors and conductors. In insulators (a) valence and conduction band are separated by a big band gap so that electrons cannot be thermally excited from the valence to the conduction band. Conductors have overlapping bands (c) or a partially filled conduction band (d) and are therefore electrically conducting.

In intrinsic (undoped) semiconductors only a small fraction of the electrons in the valence band are thermally excited into the conduction band. Extrinsic (doped) semiconductors have additional localized energy states within the band-gap. Donor states close to the conduction band (e.g. P in Si) emit their electrons into the conduction band and are (almost) completely ionized (positively charged) already well below room temperature. Acceptor states close to the valence band trap electrons and leave holes in the valence band.

In thermal equilibrium the occupation probability F of states with energy E at temperature T follows from Fermi statistics

$$F(E) = 1/\left(1 + \exp\left(E - E_{\mathrm{f}}\right)/kT\right), \tag{5.1}$$

with k the Boltzmann constant. The overall charge neutrality determines the Fermi level E_{f}.

Electrons bound in one of the localized donor states may be emitted into the conduction band by thermal excitation with a probability ε_n, thereby ionizing donors. Ionized donors may also capture electrons out of the conduction band. This process is described by a capture cross section σ_n. In thermal equilibrium these two processes have to balance each other. That condition allows to derive a relation between emission probability ε_n and capture cross section σ_n:

$$\varepsilon_n = \sigma_n \, \nu_{th\,n} \, n_i \, \exp\left((E_d - E_i)/kT\right), \tag{5.2}$$

with $\nu_{th\,n}$ thermal velocity of electrons in the conduction band, n_i intrinsic carrier concentration, E_d donor energy level, E_i intrinsic energy (Fermi level for an intrinsic semiconductor). This relation is valid more generally and can be applied to non-equilibrium conditions.

Electrons in the conduction band and holes in the valence band can move freely within the crystal lattice, their movement being only retarded by scattering on imperfections of the lattice. These imperfections may be due to lattice defects, doping atoms replacing regular atoms of the crystal (substitutional dopands) and distortions of the lattice due to thermal vibrations. The simplified way of describing these effects uses the assumption that charge carriers are accelerated by the electric field and lose all previous history at each scattering, starting with random thermal velocity again.

The movement due to the electric field is described by the drift velocity that for low fields can be assumed to be proportional to the electric field:

$$\nu_n = \left(-q\tau_c/m_n\right) E = -\mu_n \, E, \qquad \nu_p = \left(q\tau_c/m_p\right) E = \mu_p \, E, \tag{5.3}$$

with ν_n , ν_p , μ_n , μ_p being the drift velocities and low-field mobilities of electrons and holes, respectively, q elementary charge, τ_c average time between collisions, m_n , m_p effective masses of electrons and holes, and E electric field. For a high electric fields τ_c decreases and the drift velocity saturates.

At very high electric field electrons and holes may acquire sufficient energy in between collisions to generate additional electron hole pairs. This avalanche process can be the cause for an electrical breakdown of devices. It may also be used as an intrinsic amplification process in order to get sufficiently high signals from very small ionization.

For inhomogeneous carrier distributions charge carriers will preferably diffuse from high concentrations to regions of lower concentrations. This diffusion mechanism is described by

$$F_n = -D_n \, \nabla n, \qquad F_p = -D_p \, \nabla p. \tag{5.4}$$

With F_n , F_p flux of electrons and holes, D_n , D_p diffusion constant. Electron and hole current densities due to drift and diffusions are given by

$$J_n = -q\mu_n\, nE + q\, D_n\, \nabla n, \qquad J_p = q\mu_p\, pE - q\, D_p\, \nabla p. \tag{5.5}$$

Diffusion constant and mobility are related by Einstein's relation $D = (kT/q)\ \mu$. It can be derived from the requirement of zero current in thermal equilibrium of a device with non-uniform doping that has to have a constant Fermi level.

In the absence of magnetic fields charge carriers will move approximately parallel (holes) or antiparallel (electrons) to the electric field. The magnetic field adds a force perpendicular to the direction of motion and to the magnetic field direction so that the charge carriers move at an angle $\theta_p = \mu^{H}{}_p\, B$, $\theta_n = \mu^{H}{}_n\, B$ with respect to the drift direction. The Hall mobilities $\mu^{H}{}_p$ and $\mu^{H}{}_n$ differ from the drift mobilities μ_p and μ_n. B is the magnetic field component perpendicular to the electric field and the particle velocity.

5.4 Radiation Damage

Damage by ionizing and non-ionizing radiation, represents a major limitation for the use of silicon detectors in the harsh radiation environment of high-luminosity colliders, like the CERN LHC, where after its upgrade, fluences exceeding 10^{16} cm^{-2} and dose values up to 5 MGy will be reached. At high-brilliance X-ray sources, like the European X-ray Free-Electron Laser at Hamburg, dose values up to 1 GGy are expected. Radiation damage is classified in surface and bulk damage.

Surface damage is caused by ionization by charged particles and X-ray photons in the insulating layers, e.g. the SiO_2, required to fabricate silicon sensors. Like in the silicon bulk, ionizing radiation produces electron-hole pairs in the SiO_2. Whereas the mobility of electrons is sufficiently high so that they can move to a nearby electrode, holes are trapped, which results in a positive charge layer and interface traps at the Si-SiO_2 interface [4]. Positive surface charges can result in an electron accumulation layer in the Si at the interface, which can cause shorts between electrodes or break down. Interface traps, if exposed to an electric field, produce surface-generation currents. As the exact conditions at the Si-SiO_2 interface also depend on the potential on the outer SiO_2 surface, which in particular in dry conditions has a very high surface resistance (sheet resistance $> 10^{18}\ \Omega_{\square}$), it can take days until equilibrium is reached [5]. The result can be a breakdown after several days of operation or a humidity-dependent breakdown voltage. Surface radiation damage also depends on the dose-rate, which together with long time constants has to be taken into account, when studying surface damage or when testing silicon detectors. Surface damage is also technology dependent. In addition, already at room temperature significant annealing takes place. All these effects make a systematic study of surface-radiation damage difficult and time consuming. However, also thanks to the methods developed for radiation-hard electronics,

surface-radiation effects are sufficiently well understood and can be avoided by a proper design [6]. Nevertheless, there are many examples of improper designs and several unpleasant surprises due to surface damage.

Non-ionizing interactions, which knock out silicon atoms from their lattice points, are the main cause of bulk damage. A minimum energy transfer to the silicon atom of about 25 eV is required to produce such a primary defect. For energy transfers above 1 keV the silicon atom itself can knock out further silicon atoms, resulting in defect clusters, and for energies above 12 keV multiple clusters can be produced. These threshold numbers are the result of model calculations and only limited experimental information is available. The primary defects are mobile at room temperature. Some of them anneal, others diffuse to the silicon surface or interact with crystal defects and impurities and form stable defects. Using different spectroscopic methods a large number of defects could be identified and their properties, like donor- or acceptor-type, position in the band gap, cross-sections for electrons and holes and introduction rates determined [7]. The electrically active defects have three main consequences for detectors: (1) Increase of dark current, (2) trapping of signal charges thus reducing the charge collection, and (3) change of the electric field in the space charge region from which the signal charge is collected.

Typical introduction rates of stable defects are of order 1 cm^{-1}, i.e. a fluence if 1 particle per cm^2, produces 1 stable defect per cm^3. For fluences above about 10^{14} cm^{-2}, the density of defects exceeds by far the doping density, and the silicon properties change significantly: In non-depleted silicon the high generation-recombination rate results in an approximately equal density of holes and electrons, and the resistivity increases from the value determined by the dopant density to the value of intrinsic silicon, which is about 350 kΩ·cm at room temperature. The high dark current for a reverse biased diode, which is dominated by holes at the cathode and by electrons at the anode, results in a position-dependent filling of the defects and a completely different electric field distribution than in the detector before irradiation. High field regions appear at anodes and cathodes, a phenomenon called "double junction", and lower field regions in-between [8, 9]. Thus the concept of uniform doping breaks down and most of the methods used to characterize silicon before irradiation are no more applicable.

Based on a detailed and systematic study of silicon pad diodes with different doping and impurities irradiated by different particles and fluences, the phenomenological *Hamburg model* has been developed [10]. It parametrises the change of parameters like dark current and effective doping, used to characterise non-irradiated sensors, as a function of irradiation fluence and temperature history. Up to fluences of approximately 10^{14} cm^{-2}, which are presently (mid 2018) reached at the LHC, the model is remarkable successful in describing the observed effects of radiation damage. An extension of such a model to higher fluences is badly needed for monitoring the radiation fields at the LHC and for the planning of the experiments at the High-Luminosity LHC.

5.5 Semiconductor Detector Principles

The very basic and most common detector type, the reverse-biased diode, has already been sketched in Sect. 5.2. Here we will give some more information on this device and also present some more sophisticated principles, the semiconductor drift chamber and the DEPFET detector-amplification structure, while detectors based on the avalanche mechanism will be discussed in a later chapter.

5.5.1 Reverse Biased Diode (as Used in Strip and 3-D Detectors)

The principle of a reverse biased diode has already been sketched in Sect. 5.2. Here a more detailed discussion is given. Even without applying a bias the *p*-*n* junction develops a space charge region due to the diffusion of electrons and holes across the junction leading to a surplus of negative charge on the *p*-side and of positive charge on the *n*-side of the junction. This creates an electric field, a drift current and a space charge region on both sides of the junction. At any point of the device drift and diffusion currents cancel each other in equilibrium without external bias. Such a device can already be used as radiation detector since electron–hole pairs created in the space charge region will be separated by the electric field thus create a current across the junction.

Reverse biasing will increase the space charge region and therefore the electric field. For a strongly asymmetric, but in each region uniformly doped p^+n junction (as shown in Fig. 5.1) the depth of the space charge and therefore the sensitive region increases with the square root of the applied voltage.

Reverse biased diodes have been used as energy sensitive radiation detectors in Nuclear Physics for quite some time. The real breakthrough came with strip detectors in Particle Physics used for particle tracking with micro-meter accuracy. Many small strip-like diodes were integrated on the same wafer and each one connected to its own readout channel (Fig. 5.3). The particle position was given by the channel giving the signal. More sophisticated strip detectors will be described in Sect. 5.6.

Planar pixel detectors are obtained by shortening the individual strips so that they do not reach anymore the detector edge and form a two-dimensional pattern. Detectors with pixel sizes down to $15 \times 15\ \mu m^2$ have been built. The main difficulty of such detectors is their readout. Different realisations will be discussed later.

A different concept of diode detectors, the so called 3-D detectors [11], is shown in Fig. 5.4: Holes with diameters of a few micro-meters are etched into the crystal orthogonal to its surface, and alternate holes are n^+- and p^+-doped. A voltage difference between the n^+- and p^+-doped columns generates an electric field parallel to the crystal surface. The number of electron-hole pairs produced by a charged particle traversing the detector at large angles to the surface is given by

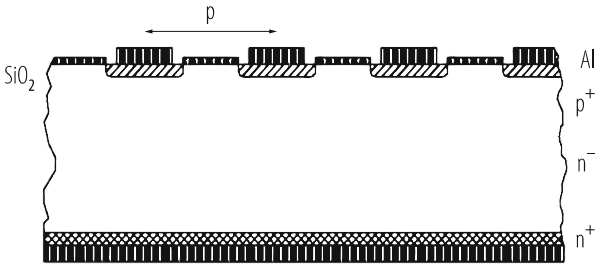


Fig. 5.3 Cross section of a silicon strip detector built on lightly phosphor doped (n^-) silicon bulk material. Strips are highly boron doped (p^+) and the backside highly phosphor doped (n^+)

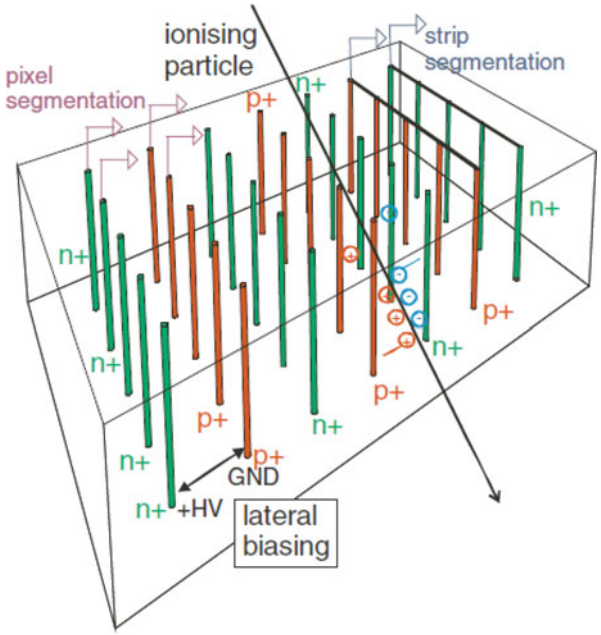


Fig. 5.4 Principle of the 3-D detector: Holes are etched into the silicon crystal orthogonal to the detector surface. Alternate holes are n^+- and p^+-doped. A positive voltage on the n^+-contacts, with the p^+-contacts grounded, generates an electric field parallel to the detector surface. In a 3-D detector the charge generated, given by the crystal thickness, and the charge collection distance, given by the distance between the holes, can be separately chosen (Book F. Hartmann Fig. 1.69)

the crystal thickness, whereas the charge collection distance is given by the column distance. In this way signal and charge collection distances can be chosen separately and the detector can be optimised for radiation tolerance. In addition, the operating voltage for 3D-detectors and thus the power heating the detector are significantly reduced compared to planar detectors. By connecting the p^+- and n^+-columns with

different metal patterns, strip- and pixel-sensors and other readout geometries can be realized.

5.5.2 Semiconductor Drift Chamber

The semiconductor drift chamber has been invented by Emilio Gatti and Pavel Rehak [1]. This device (Fig. 5.5) makes use of the sideward depletion principle, having diode junctions on both surfaces and a bulk contact on the fringe. Fully depleting the device by applying a reverse bias voltage between *p*- and *n*-contacts creates a potential valley for electrons in the middle plane. Electrons created by ionizing radiation will assemble in this valley and subsequently diffuse until they eventually reach the *n*-doped anode. Faster and controlled collection is achieved by adding a horizontal drift field. This is obtained by dividing the diodes into strips and applying from strip to strip increasing voltages.

This device is able to measure position (by means of the time difference between particle interaction and arrival of the signal at the anode) as well as the energy from the amount of signal charge. In many applications the latter aspect is the important one. Here one profits from the small electric capacitance of the anode compared to the planar diode shown in Fig. 5.1, which acts as capacitive load to the readout amplifier. Large area detectors can therefore be operated with excellent energy resolution at high rates.

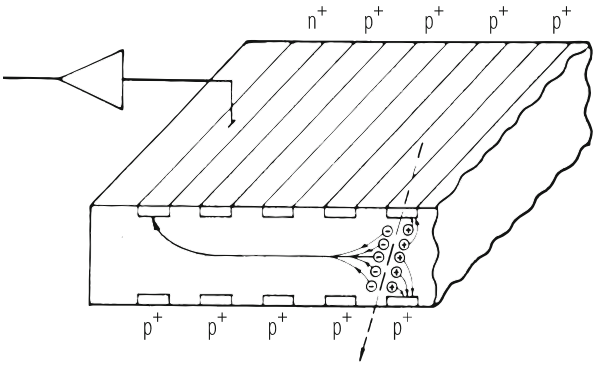


Fig. 5.5 Semiconductor drift chamber using the sideward depletion method. Dividing the p^+ doped diodes into strips and applying a potential which increases from strip to strip superimposes a horizontal field in the potential valley that drives the electrons towards the n^+ anode which is connected to the readout electronics. Upon arrival of the signal charge at the n^+ anode the amount of charge and the arrival time can be measured

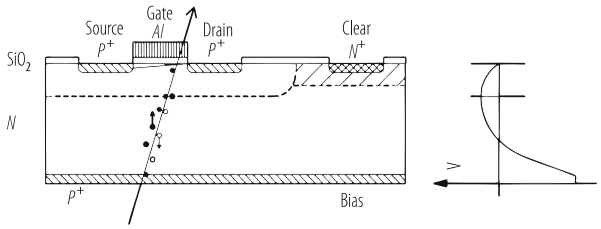


Fig. 5.6 The concept of a DEPFET: Simplified device structure (*left*) and potential distribution along a cut across the wafer in the gate region of the transistor (*right*)

5.5.3 DEPFET Detector-Amplification Structure

The *DEPFET* structure which simultaneously possesses detector and amplification properties has been proposed by J. Kemmer and G. Lutz in 1987 [3] and has subsequently been confirmed experimentally [12]. It is based on the combination of the sideward depletion method—as used in a semiconductor drift chamber shown in Fig. 5.5—and the field effect transistor principle.

In Fig. 5.6 a *p*-channel transistor is located on a fully depleted *n*-type bulk. Compared to Fig. 5.5 the potential valley has been moved close to the top side. Signal electrons generated in the fully depleted bulk assemble in a potential minimum for electrons ("internal gate") and increase the transistor channel conductivity in a similar way as by changing the (external) gate voltage. The device can be reset by applying a large positive voltage on the clear electrode.

The DEPFET has several interesting properties:

- Combined function of sensor and amplifier;
- Full sensitivity over the complete wafer;
- Low capacitance and low noise;
- Non-destructive repeated readout;
- Complete clearing of the signal charge: No reset noise.

These properties make it an ideal building block for an X-ray pixel detector, or for a pixel detector for the precision tracking of charged particles.

5.6 Silicon Strip Detectors (Used in Tracking)

Silicon strip and pixel detectors are the most common semiconductor detectors in Particle Physics, mostly used for particle tracking. There one profits from the precise position measurement (few μm) at very data rates (up to tenths of MHz per detection element). In its simplest form they are narrow strip diodes put next to each other on the same semiconductor substrate, each strip having its own readout channel. Typical charge collection times are about 10 ns. Due to diffusion, track

inclination and the Lorenz force in a magnetic field, the charge of one track may be distributed over two or more strips. This can be exploited to improve the accuracy of the position measurement well below the value given by the strip pitch. It is then limited by fluctuations of the ionization process and in particular the generation of delta electrons and the electronics noise. A measurement precision down to about 1 μm has been achieved.

5.6.1 Strip Detector Readout

In the conceptually simplest version each strip is connected to its own electronic readout channel and the position is determined by the number of the strip providing a signal.

Binary (yes/no) *readout* may be used if no energy information is required and if the position accuracy given by the strip pitch is sufficient. One also does not lose position resolution compared with analogue readout if the strip pitch is large with respect to the width of the diffusion cloud.

Analogue (signal amplitude) *readout* of every channel may lead to a substantial improvement of the position measurement precision if the strip spacing matches the charge spread due to diffusion during collection. (Charge spread can also be due to track inclination or the Lorenz angle in a magnetic field.) In addition, the simultaneous measurement of energy loss becomes possible.

Charge division readout reduces the number of readout channels as only a fraction of the strips is connected to a readout amplifier (Fig. 5.7). Charge collected at the other (interpolation) strips is divided between the two neighbouring readout channels according to the relative position. Charge division is due to the capacitors between neighbouring strips. For charge division to work, it is necessary to hold the intermediate strips at the same potential as the readout strips. This can be accomplished by adding high ohmic resistors or with other methods. If the intermediate strips were left floating, they would adjust themselves to a potential

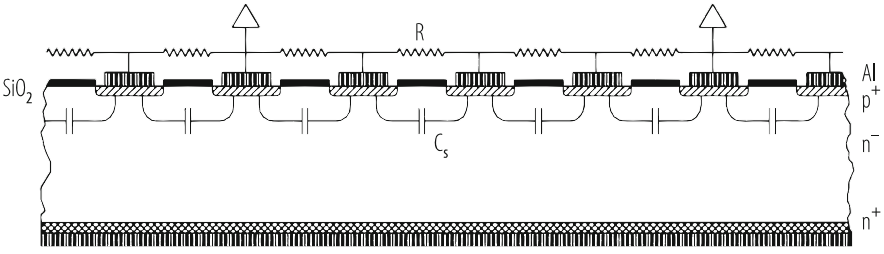


Fig. 5.7 Charge division readout. The interstrip capacitors between the readout strips act as capacitive charge divider. The high-ohmic resistors are required to keep all strips at the same potential

such that they would collect no signal charge, and thus charge division would cease to function.

5.6.2 *Strip Detectors with Double-Sided Readout*

As shown in Fig. 5.8 it is possible to segment the electrodes on both sides of the wafer. This double-sided readout has the obvious advantage of providing twice the information for the same amount of scattering material. With crossed strips on the two detector faces, a projective two-dimensional measurement is obtained from a single detector.

For a traversing particle, a spatial point can be reconstructed as both projections are obtained from the same initial charge cloud. With analogue readout it is furthermore possible (to some degree) to correlate signals from the two sides, making use of Landau fluctuations and the equality of the charge induced on both sides for each ionizing particle. This can be of interest for resolving ambiguities when several particles traverse simultaneously the detector.

A problem in producing double-sided detectors is the insulation of neighbouring strips on both detector sides. The naive solution of only providing highly doped *n*- and *p*-doped strips on the two sides of the detector (Fig. 5.8) fails because of the build-up of an electron-accumulation layer (an inversion layer on *p*-type silicon) between the *n*-strips below the insulating SiO_2 (Fig. 5.9a). This electron layer results in an electrical shortening of neighbouring strips. It is caused by the positive charges that are always present at the silicon-oxide interface. As discussed in Sect. 5.4 ionizing radiation results in a further increase of positive charges.

There are three possibilities for curing the problem:

1. Large-area *p*-type surface doping. In this case the oxide charges are compensated by the negative acceptor ions and the build-up of the electron layer is prevented (Fig. 5.9b). This method requires a delicate choice of *p*-type doping concentration and profile. A too large doping results in high electric fields and in a possible

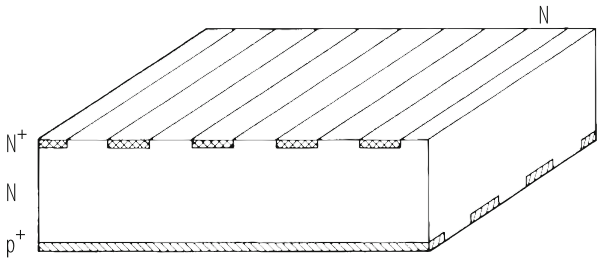


Fig. 5.8 Double sided strip detector (naive solution)

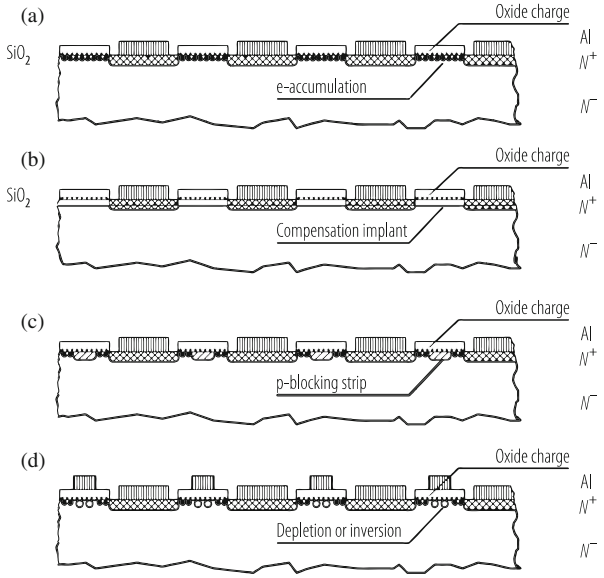


Fig. 5.9 Insulation problem for *n*-strips in silicon, due to electrical shortening by an electron accumulation layer (**a**), and three possible solutions: Large area *p*-implantation (**b**); interleaved *p*-strips (**c**) and negatively biased MOS structures (**d**)

electrical breakdown at the strip edges. This problem is alleviated by the other two solutions presented below.

2. Disruption of the electron layer by implantation of *p*-strips between the *n*-doped charge-collection strips (Fig. 5.9c); and
3. Disruption of the electron layer by a suitably biased (negatively with respect to the *n*-strips) MOS structure (Fig. 5.9d). For moderate biasing neither electrons nor holes will accumulate underneath the MOS structure, while for a high negative bias a hole layer (inversion on *n*-type, accumulation on *p*-type silicon) will form.

5.6.3 Strip Detectors with Integrated Capacitive Readout Coupling and Strip Biasing

Capacitive-coupled (AC) readout (Fig. 5.10, *right*) has the obvious advantage of shielding the electronics from dark current, whereas direct coupling (DC, Fig. 5.10, *left*) can lead to pedestal shifts, a reduction of the dynamic range, drive the electronics into saturation or requires a dark-current compensation.

As it is difficult to fabricate high-ohmic resistors, and almost impossible to produce sufficiently large capacitors in LSI electronics, it seemed natural to integrate

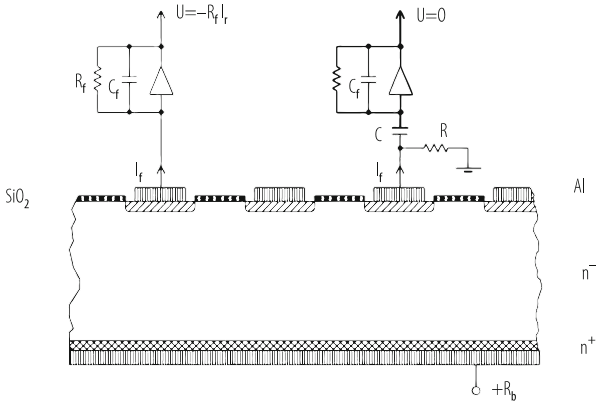


Fig. 5.10 Direct and capacitive coupling of electronics to the detector. With direct coupling (*left*) the detector reverse bias current I_f has to be absorbed by the electronics. With capacitive coupling (*right*), only the AC part of the detector current reaches the electronics, while the DC part flows through the resistor R

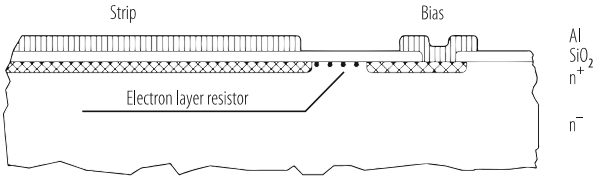


Fig. 5.11 *n*-strip biasing by an electron-accumulation-layer resistor. The diagram shows a cut along the strip direction. The electron layer is induced by the always present positive oxide charges that attract electrons towards the Si-SiO_2 interface. It is sidewise enclosed by p-implants so as to prevent electrical shortening between neighbouring strips. Bias and strip implants are at nearly the same potential

these elements into the detector. This has been done in a collaborative effort by a CERN group with the Center of Industrial Research in Oslo [13], where the detectors were produced. Capacitances have been built by separating implantation and metallization of the strips by a thin SiO_2 layer. Biasing resistors were made of lightly doped polysilicon, a technology that is used in microelectronics. The detectors gave very satisfactory results. The strip detectors of several particle physics experiments use this design.

A different method of supplying the bias voltage to the detector has been developed and used for double-sided readout by a Munich group [3, 12]. It leads to a considerable simplification of the technology as it does not require resistors but only uses technological steps that are already required for DC coupled detectors. The polysilicon technology can be avoided altogether; instead, the voltage is supplied through the silicon bulk. Two methods can be applied either using the resistance of an electron accumulation layer (Fig. 5.11) that is induced by the positive oxide charge or a punch through mechanism that occurs between two closely spaced *p*-

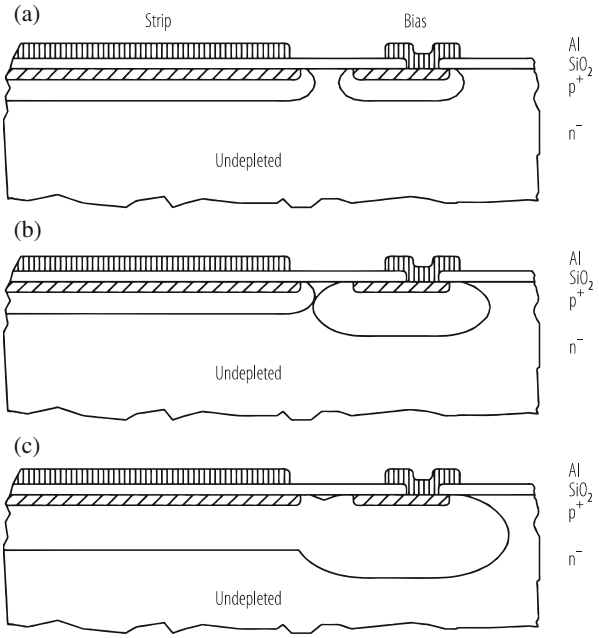


Fig. 5.12 *p*-strip punch-through biasing. The diagrams show cuts along the strip direction: (**a**) Before applying a bias voltage, where the space-charge regions around the strip and the bias implant are isolated; (**b**) at onset of punch-through, where the space-charge region around the bias implant has grown and just touches the space-charge region of the strip. The potential barrier between strip and bias implants has diminished, but is just large enough to prevent the thermal emission of holes towards the bias strip; (**c**) at larger bias voltage, where the space-charge region has grown deeper into the bulk. Holes generated in the space-charge region and collected at the strip implant are thermally emitted towards the bias strip. The voltage difference between strip implant and bias depends on geometry, doping and bias voltage. A weak dependence on oxide charge is also present

electrodes (Fig. 5.12). These biasing methods can be used for single sided and also for double sided readout where *p*- and *n*-strips are located at opposite surfaces of the wafer as was the case in the ALEPH experiment. In all cases the capacitors are built by interleaving a thin oxide layer between implantation and metal strips.

A word of caution on the operation of capacitive-coupled detectors and in particular of double sided detectors will be given at this point since it has been overlooked in a couple of experiments causing detector breakdown. At first glance it seems that one can choose the voltages on implant and metal strips independently. However this can result in shortening of neighbouring strips or electrical breakdown due to the build-up of accumulation layers at the Si-SiO_2 interface. Although the SiO_2 is not covered with an ohmic layer its surface will slowly charge up to a potential close to the neighbouring metal electrodes, because of a high but finite surface resistivity, as discussed in Sect. 5.4.

5.7 Detector Front-End Electronics

Before discussing more sophisticated detectors we now turn to readout electronics, a subject relevant to all detectors. As there is a close interplay between a detector and its electronics, both components have to be considered together when designing a detector for a specific application. In most cases a signal charge produced by photons or ionizing radiation has to be measured as precisely as possible in a predefined time interval and with tolerable power consumption. Readout uses in most cases large scale integrated (LSI) electronics adapted to the needs of the special application.

5.7.1 Operating Principles of Transistors

Transistors are commonly classified into unipolar and bipolar, depending on whether only one or both types of charge carriers participate in the current flow. As a consequence of the difference in operating principles, their properties—and therefore their suitability for specific applications—differ greatly. Bipolar transistors are well suited for high-speed applications and for driving large currents. Unipolar transistors are common in moderate-speed low-noise applications (JFETs) and are most prominent in digital circuitry (MOSFETs).

We use as an example the *n*-channel MOSFET (Metal-Oxide-Semiconductor Field Effect Transistor). Figure 5.13 shows a cross section along the channel. Two n^+p diodes are connected by a MOS structure. Applying a high enough positive potential on the gate an inversion (electron) layer will connect source and drain and for non-zero drain-source voltage an electron current will flow from source to drain. The strength of this current can be controlled by the gate potential and also

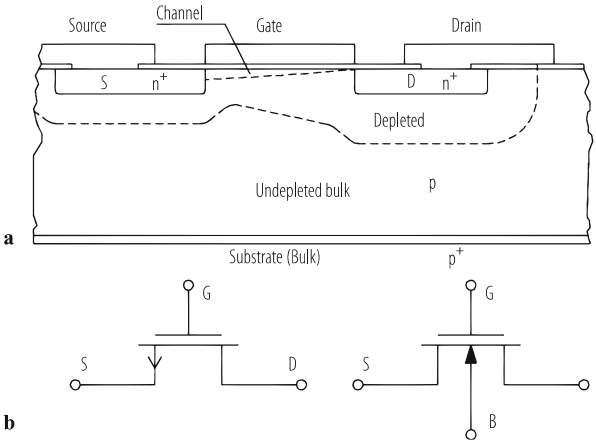


Fig. 5.13 *n*-channel MOSFET: Cross-section (**a**) and device symbol (**b**). The separation of the space-charge region from the channel below the gate and from the undepleted bulk is indicated by the dashed lines

by the drain voltage. A resistive voltage drop along the channel is responsible for the current saturation that occurs once this voltage drop equals the effective gate voltage (voltage above the threshold necessary to create inversion).

Important parameters of the transistor to be used in noise considerations are the transistor (output) conductance $g = \mathrm{d}I_\mathrm{d}/\mathrm{d}V_\mathrm{d}$ and transconductance $g_\mathrm{m} = \mathrm{d}I_\mathrm{d}/\mathrm{d}V_\mathrm{g}$. These and other parameters can be modelled using the graded channel approximation which relies on the assumption that changes along the channel are much smaller than those occurring in the transverse direction. It allows deriving scaling laws for changes in geometry. However, for microelectronics with minimal feature size these are of limited validity. Instead, two- and three-dimensional numerical device simulations are needed.

Measurement precision is limited by noise. There are several noise mechanisms present. Considering a resistor with resistance R for example, the thermal motion of electrons will result in a statistical fluctuation of the charge distribution in the conductor, leading to a noise voltage density of $\mathrm{d}\langle v_\mathrm{n}^2\rangle/\mathrm{d}f = 4\ kT{\cdot}R$ between the terminals of the resistor. The resistance of the MOSFET channel is a source of white noise too. It is customary to represent this noise by a voltage at the gate $\mathrm{d}\langle v_\mathrm{n}^2\rangle/\mathrm{d}f = 4\ kT(2/3)(1/g_\mathrm{m})$ for the operation of the transistor in the saturation region.

A further mechanism of noise is the capture and delayed release of single charge carriers in the channel. While being captured the drain current decreases, returning to the initial value when released. For a single trapping centre with characteristic average capture and release times a Lorentzian noise spectrum as function of frequency results. Having many different trapping centres, as is the case for traps at the $Si\text{-}SiO_2$ interface where trapping and detrapping occurs by means of tunnelling, the result of the superposition of Lorentzian noise spectra is a $1/f$ spectrum $\mathrm{d}v_\mathrm{n}^2/\mathrm{d}f = A_f/f$ with A_f a constant, which depends on the technology and the geometric parameters of the transistor. A_f is usually obtained from measurements and parameterized as $A_f = K_\mathrm{F}/(WLC_\mathrm{ox}^2)$. K_F characterizes the technology, W and L are channel width and length, and C_ox the oxide capacitance per unit area. Note that the *1/f* noise is independent of the transistor current.

5.7.2 The Measurement of Charge

The standard problem in the readout of a semiconductor detector is the low-noise measurement of the signal charge, usually under severe constraints such as high-speed operation, low power consumption, restricted space and frequently high radiation levels. In this section the general problems of charge measurement will be addressed, while specific solutions for the electronics will be considered later.

The *charge-sensitive amplifier* (CSA), invented by Emilio Gatti [14] and represented in Fig. 5.14, consists of an inverting amplifying circuit which—in the ideal case—delivers an output voltage proportional to the input ($U_\mathrm{out} = -A\ U_\mathrm{in}$) and a feedback capacitor C_f. In addition, a high-resistance feedback or a switch is needed in the feedback loop, in order to bring the circuit into its operating condition. C_D

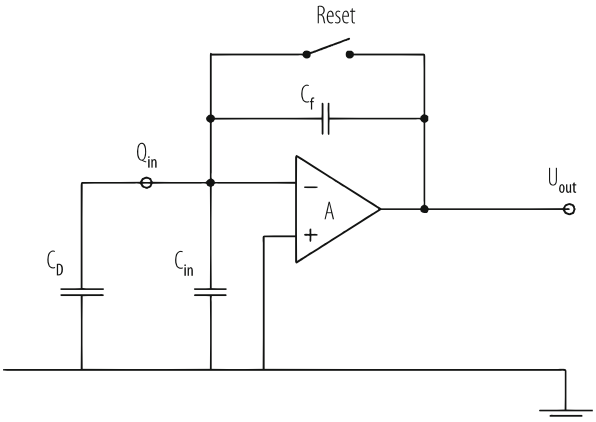


Fig. 5.14 Principle of a Charge Sensitive Amplifier (CSA). The inverting amplifier has gain A and a capacitive feedback. The reset switch is only used for bringing the system into its operating condition, and is often replaced by a high-ohmic resistor

represents the capacitive load of the detector at the input, C_{in} the capacitive load to ground in the amplifier, which is usually dominated by the gate capacitance of the input transistor.

Putting a charge Q_{in} at the input will result in an output voltage change of $U_{\mathrm{out}} = -Q_{\mathrm{in}} /(C_{\mathrm{f}} +(C_{\mathrm{D}}+C_{\mathrm{in}}+C_{\mathrm{f}})/A)$ which for large amplification is given by the ratio of signal charge over feed-back capacitance, indicating that the charge has been transferred completely from the detector to the feedback capacitor. For low frequencies the input impedance of the CSA will be represented by a capacitance of the value $C_{\mathrm{eff}} = (A+1)\ C_{\mathrm{f}} + C_{\mathrm{in}}$. A high value of $C_{\mathrm{eff}} > C_{\mathrm{D}}$, i.e. a low input impedance, is important because when C_{eff} is only of the same order of magnitude as the detector capacitance C_{D} the charge is incompletely transferred to the electronics. This results in a loss of sensitivity and possibly crosstalk within the detector to neighbouring channels.

Turning now to the question of measurement precision, respectively noise in the detector-amplifier system, we remark that it is customary to represent the effect of all amplifier noise sources by a single noise voltage U_{n} placed at the input (Fig. 5.15). As this noise voltage generator is in series with detector and amplifier it is called serial noise. The presence of the serial noise voltage U_{n} will result in an output voltage even if there is no signal charge present. For an evaluation of the serial noise charge, it is easiest to consider the charge necessary to compensate for the effect of the noise voltage, such that the output voltage remains at zero. The value can be immediately read from Fig. 5.15: $Q_{\mathrm{n}} = U_{\mathrm{n}}\ (C_{\mathrm{D}} + C_{\mathrm{in}} + C_{\mathrm{f}}\ ) = C_{\mathrm{T}} U_{\mathrm{n}}$ with C_{T} the total "cold" input capacitance.

Notice that the serial noise is generated in the amplifier, the influence of the detector is due to the capacitive load at the amplifier input only. The detector itself produces noise due to statistical fluctuations of its leakage current I. This parallel

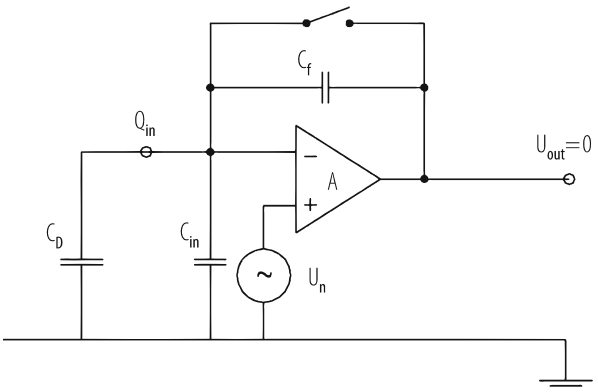


Fig. 5.15 The effect of amplifier serial noise in a detector-amplifier system

noise is represented by a noise current source of density $d<i_n^2>/df = 2I \cdot q$ in parallel to the detector capacitance C_D. To estimate the charge measuring precision one has to follow separately signal and noise through the complete readout chain and compare their respective output signals.

The signal produced by the amplifier will usually not be used directly; it will be further amplified and shaped, in order to optimize the ratio of signal to noise and to reduce the interference between subsequent signals. We will only consider a few very simple cases, the simplest being an idealized charge-sensitive amplifier followed by an RCCR filter. For a more elaborate treatment, the reader is referred to the literature (e.g. [15]).

The arrangement of a CSA followed by an RCCR filter is shown in Fig. 5.16. The output of the CSA is a voltage step given for very high amplification as Q/C_f. The shaper does an RC integration followed by a CR differentiation. This procedure results in a signal peak, which for the same integration and differentiation time constant $\tau = R_1C_1 = R_2C_2$ has the shape $U_{out}(t) = (Q/C_f) \cdot (t/\tau) \cdot \exp(-t/\tau)$ with a peak value $U_{peak} = (Q/C_f) \cdot \exp(-1)$. The height of this peak is a measure of the signal charge. Superimposed on the signal is the noise voltage, and we are interested in the signal-to-noise ratio, which is defined as the ratio of the height of the peak value to the root-mean-square value of the noise voltage measured at the same point in the circuit.

In order to find the noise voltage at the output, each noise source in the circuit has to be traced to the output and the resulting voltages added in quadrature. Doing so, one finds the important result that, for white (thermal) serial noise, the ratio of noise to signal (N/S) decreases with the square root of the shaping time constant τ, while for *1/f* noise this ratio remains constant. Parallel noise, given as a time integral over current fluctuations, increases with the square root of the shaping time.

More sophisticated continuous time filtering methods use (for example) Gaussian shape filtering, which can be approximated by several sequential RC integration and differentiation steps. Especially important in integrated electronics are the

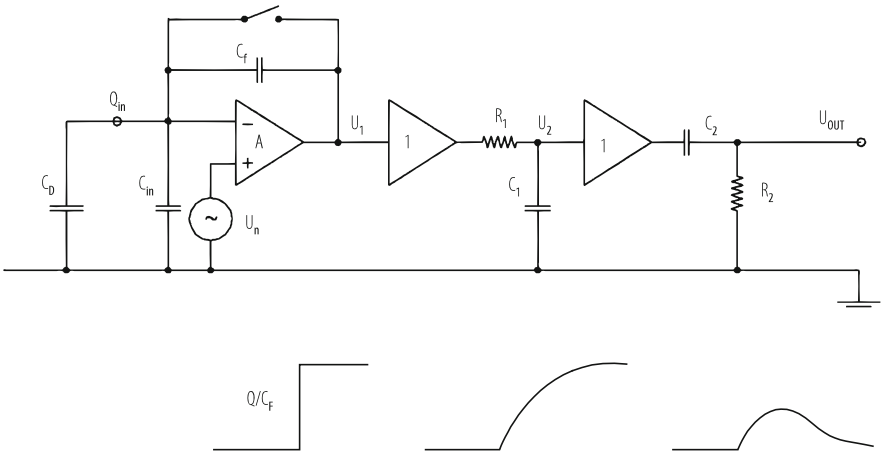


Fig. 5.16 Noise filtering and signal shaping in an RCCR filter following a charge-sensitive amplifier (top). The two unity gain amplifiers have been introduced in order to completely decouple the functions of the CSA, the integration (RC) and the differentiation (CR) stages. The signal form is indicated for each stage (bottom)

techniques in which the output signal is sampled several times and mathematical manipulations of the samples are performed. This can be done either after the measurement by numerical processing or directly by the local readout electronics. In the latter case, it is usually achieved by using switched capacitor techniques for analogue algebraic manipulations. Common to both methods, however, is the need to sample the signal at fixed (or, at least, known) times with respect to its generation. Alternatively with frequent enough sampling, the arrival time of the signal can be extracted from the data and filtering can be done afterwards by selecting the relevant samples before and after arrival of the signal. In all cases, however, the fact that the three noise components (white serial, 1/*f* and white parallel noise) scale with the available readout time in the described way remains valid.

As a further example we discuss double correlating sampling realized in switched capacitor technology which is most naturally realizable in integrated circuit technology. It is applicable if the signal arrival time is known in advance, as is the case for example in collider physics experiments.

The circuit (Fig. 5.17) consists of two sequential charge-sensitive amplifiers connected by a coupling capacitor C_s and switch S_c. Initially all switches are closed. Thus both CSAs have reset their input and output voltages to proper working conditions and a possible offset voltage between CSA1 and CSA2 is stored on capacitor C_s. The following operations are performed in sequence: (1) opening switch S_1 at time t_1, resulting in an unwanted charge injection into the input of CSA1 and therefore an output voltage change that will be stored on capacitance C_s and thus made invisible to the input of CSA2; (2) opening of reset switch S_2. Any voltage change on the output of CSA1 (e.g. signal or noise) is also seen in the output of CSA2, amplified by the ratio C_s/C_{f2}; (3) signal charge Q_s generation at time t_3

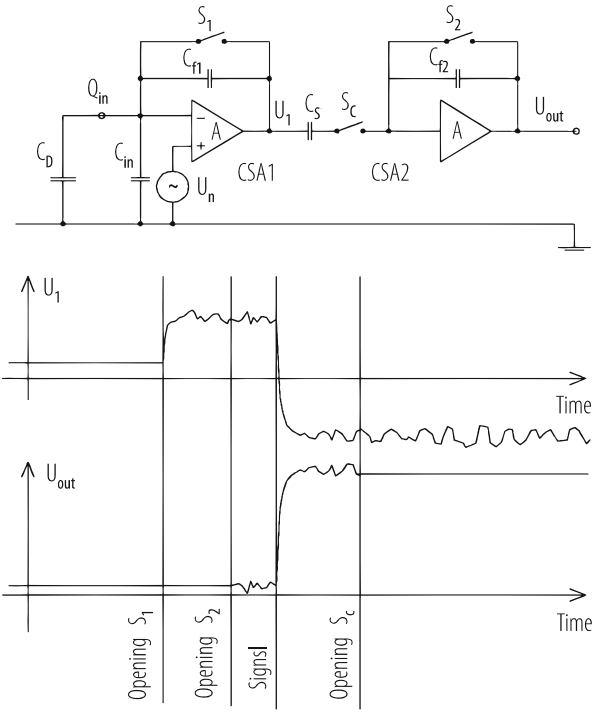


Fig. 5.17 Double-correlated sampling of the output of a charge-sensitive amplifier (CSA1) with the help of a coupling capacitor C_s and a second CSA2

changes the output of CSA1 by $\Delta U_1 = Q_s/C_{f1}$ and the output voltage by $\Delta U_{out} = Q_s\ (C_s\ /(C_{f1}C_{f2}))$; and (4) opening of switch S_c at time t_4 inhibits further change of the output voltage. The difference of the output voltage of CSA1 (amplified by C_{f2}/C_s) between times t_2 and t_4 remains present at the output of the circuit.

Double correlated sampling suppresses the reset noise due to operating switch S_1 and also suppresses low frequency noise but enhances the noise at higher frequencies. As a result white noise is not suppressed. This has to be done by limiting the frequency range of the amplifier. More sophisticated schemes of switched capacitor filtering, taking several samples (sometimes with different weights), have also been implemented.

5.7.3 Integrated Circuits for Strip Detectors

The development of integrated detector readout electronics was initiated by the simultaneous requirements of high density, low power and low noise for use with silicon strip detectors in the tight space environment of elementary particle physics

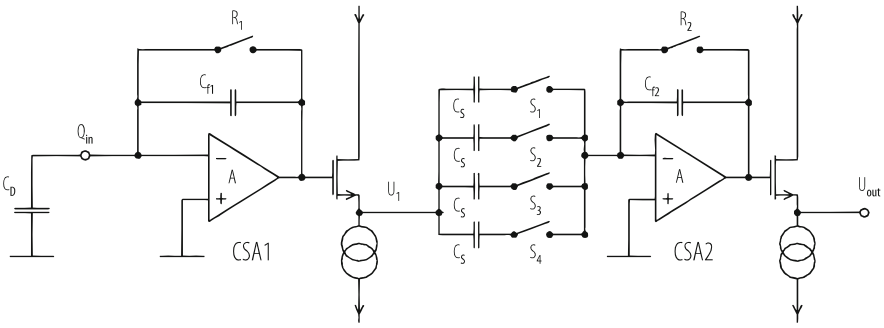


Fig. 5.18 Single channel readout schematics of the CAMEX64 strip-detector readout circuit

collider experiments. A variety of circuits has been developed for this purpose, the basic principle of essentially all of them being: (1) parallel amplification using a charge-sensitive amplifier at each input; (2) parallel signal filtering combined with second-stage amplification and parallel storage within capacitive hold circuits; and (3) serial readout through one single output channel.

We will present only one of the developments [16]. This was not only one of the first to be started but is still in use and has been further developed for many important applications.

The basic functional principle of a single channel is shown in Fig. 5.18. It consists of two charge-sensitive amplifiers, each followed by a source follower, and four sets of capacitors and switches that connect the output of the first amplifier with the input of the second amplifier. The circuit is rather similar to the one shown in Fig. 5.17, but the essential difference is the fourfold multiplication of the capacitive coupling between the amplifiers. In this way it is possible to perform fourfold double-correlated sampling at times that are shifted relative to each other. This procedure provides a good approximation to trapezoidal shaping, which means averaging the output over time intervals before and after signal arrival and taking the difference between the averaged samples.

The switching sequence that performs this function is the following: (1) Close R_1 and R_2. The charges on the feedback capacitances C_{f1} and C_{f2} are cleared. (2) Open R_1: some (unwanted) charge will be injected into the input by the switching procedure, producing an offset in U_1. (3) Close and open in sequence S_1 to S_4. The U_1 offset values at the four times t_1 to t_4 will be stored on the four capacitors C_s. (4) Open switch R_2. A small offset voltage appears at the output. (5) Deposit signal charge Q_{sig} at input. U_1 changes by an amount of $\Delta U_1 = Q_{sig}/C_{f1}$. (6) Close and open S_1 to S_4 in sequence at times t_1 to t_4. A charge $C_s \Delta U_1$ is inserted into the second amplifier at each sample. The total output voltage is $4C_s \Delta U_1/C_{f2} = Q_{sig} 4C_s/(C_{f1} C_{f2})$.

The complete chip, containing 64 channels, also comprises additional electronics, as shown in Fig. 5.19. Three test inputs allow injection of a defined charge through test capacitors. Digital steering signals are regenerated by comparators. The

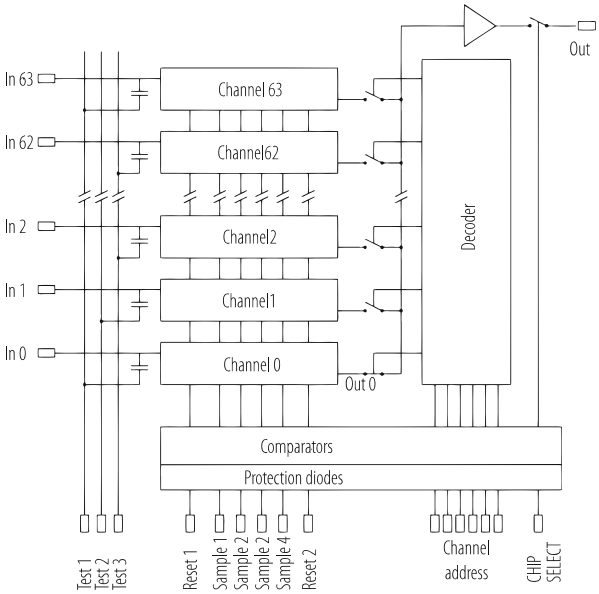


Fig. 5.19 Block diagram of the CAMEX64 strip-detector readout chip

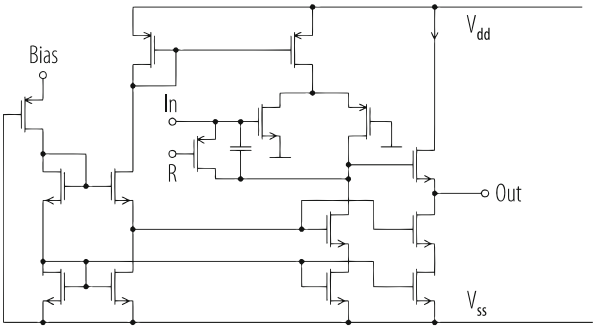


Fig. 5.20 Circuit diagram of the amplifier, including source follower and biasing circuit, of the CAMEX64 strip-detector readout chip

decoder switches one signal at a time on the single output line where a driving circuit for the external load is attached.

A circuit diagram valid for all charge sensitive amplifiers used is shown in Fig. 5.20. The current in all transistors can be scaled by a reference bias current (Bias). The input (In) can be shorted to the output with the reset switch (R) that lies in parallel to the feedback capacitor. The CSA output is connected to a source follower driving the output node (Out).

The circuits mentioned so far have been designed for moderate speed of applications in low-radiation environments. For the CERN Large Hadron Collider

(LHC), where the time difference between consecutive crossings of particle bunches (25 ns) is much shorter than the time it takes to decide whether or not the data of a particular event needs to be kept (approximately 2 μs), chips with high-speed operation and radiation hardness have been developed successfully. In addition to fast low-noise amplifiers and radiation hardness, it is required to store the information for approximately hundred bunch crossings.

The task of designing radiation hard electronics has been considerably eased by the industrial development of submicron integrated circuit technology which, due to the use of ultra-thin oxide, to a large extend has eliminated the problem of radiation induced threshold shifts in MOS transistors [17]. Taking some precautions in the design these technologies can be considered "intrinsically radiation hard".

5.8 Silicon Drift Detectors

The semiconductor drift detector was invented by E. Gatti and P. Rehak [1]. First satisfactorily working devices in silicon were realised in a collaborative effort by J. Kemmer at the Technical University Munich, the Max Planck Institute for Physics in Munich and the inventors [18].

The working principle may be explained by starting from the diode (Figs. 5.1 and 5.21a) if one realizes that the ohmic n^+ contact does not have to extend over the full area of one wafer side but can instead be placed anywhere on the undepleted conducting bulk (Fig. 5.21b). Then there is space to put diodes on both sides of the wafer (Fig. 5.21c). At small voltages applied to the n^+ electrode, there are two space-charge regions separated by the conducting undepleted bulk region (hatched in Fig. 5.21). At sufficiently high voltages (Fig. 5.21d) the two space-charge regions will touch each other and the conductive bulk region will retract towards the vicinity of the n^+ electrode. Thus it is possible to obtain a potential valley for electrons in which thermally or otherwise generated electrons assemble and move by diffusion only, until they eventually reach the n^+ electrode (anode), while holes are drifting rapidly in the electric field towards the p^+ electrodes.

Based on this double-diode structure the concept of the drift detector is realised by adding an additional electric field component parallel to the surface of the wafer in order to provide for a drift of electrons in the valley towards the anode. This can be accomplished by dividing the diodes into strips and applying a graded potential to these strips on both sides of the wafer (Fig. 5.5).

Other drift field configurations (e.g. radial drift) can be obtained by suitable shapes of the electrodes. Drift chambers may be used for position and/or energy measurement of ionizing radiation. In the first case the position is determined from the drift time. Furthermore, segmenting the n^+-strip anode in Fig. 5.5 into pads, a two-dimensional position measurement is achieved.

Due to the small capacitive load of the readout electrode to the readout amplifier, drift detectors are well suited for high precision energy measurement.

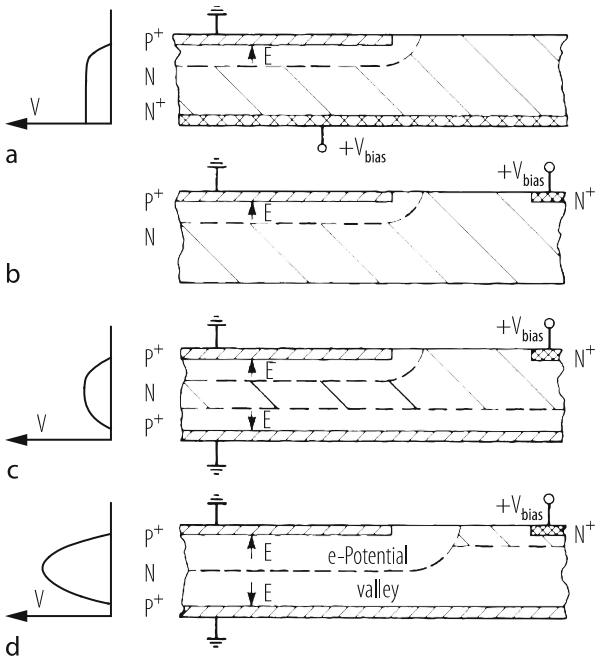


Fig. 5.21 Basic structures leading towards the drift detector: diode partially depleted (**a**); diode with depletion from the side (**b**); double diode partially depleted (**c**); double diode completely depleted (**d**)

5.8.1 Linear Drift Devices

Although linear devices seem to be the most straightforward realisation of the drift detector principle, one encounters some nontrivial problems. They are due to the finite length of the biasing strips and the increasing potential to be applied to these strips, which leads to a very large voltage of several hundred or (for very large drift length) a few thousand volts. Therefore guard structures have to be implemented which provide a controlled transition from the high voltage to the non-depleted region at the edges of the device.

A schematic drawing of the first operational silicon drift detector [18] is shown in Fig. 5.22. Anodes placed on the left and right side of the drift region collect the signal electrons generated by the ionizing radiation. The most negative potential is applied to the field-shaping electrode in the centre. Electrons created to the left (right) of this electrode will drift to the left (right) anode. The p^+-doped field electrodes do not simply end on the side, but some of them are connected to the symmetrical strip on the other half of the detector. In this way one insures that the high negative potential of the field strips drops in a controlled manner towards the potential of the undepleted bulk on the rim of the detector.

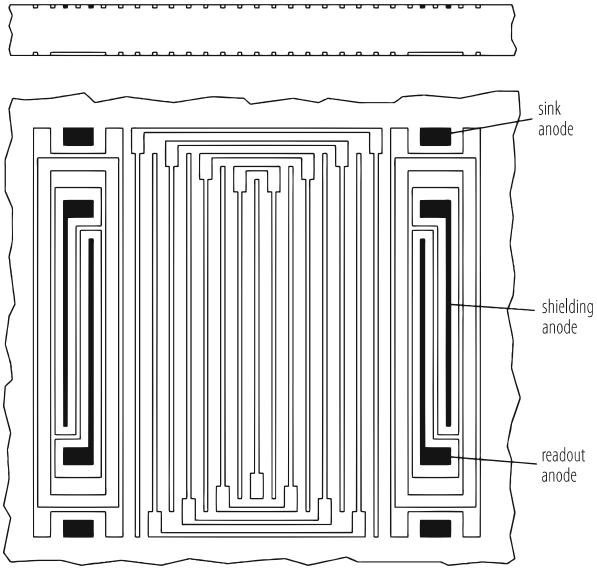


Fig. 5.22 Schematic cross-section and top view of a linear drift detector with *p*-doped field-shaping electrodes (light) and two *n*-doped (double) anodes (dark)

Looking closely at the anodes (Fig. 5.22), it can be seen that there are pairs of *n*-doped strips. Each pair is surrounded by a *p*-doped ring, which also functions as the field-shaping electrode closest to the anode. The two *n*-doped strips are separated by a *p*-doped strip that also connects to the ring surrounding the anode. Surrounding the *n*-strips completely by *p*-doped regions ensures that the adjacent *n*-doped anodes are electrically disconnected to each other and to the other regions of the detector (such as the non-depleted bulk). The outer *n*-strips are used to drain away electrons from the high voltage protection region, while the inner strips measure the signals created in the active detector region.

The opposite side of the silicon wafer is for the large part identically structured. Differences are only in the anode region, where the *n*-implantation is replaced by *p*-doped strips. In the main part of the detector, the strips on opposite sides of the wafers are kept at the same potential, thus assuring a symmetrical parabolic potential distribution across the wafer (Fig. 5.23a). Near the anode an increasing potential difference between the two wafer surfaces moves the potential valley for electrons to the front side until it ends at the anode (Fig. 5.23b).

The linear drift detectors described so far allow one dimensional position measurement only. Dividing the anode of a linear drift detector into pads (Fig. 5.24) leads to a two-dimensional position measurement. One coordinate is obtained from the drift time, the other from the pads on which the signals appear. The second coordinate may be further improved by interpolation using the signal in neighbouring pads. The signal will be distributed over more than one pad if the

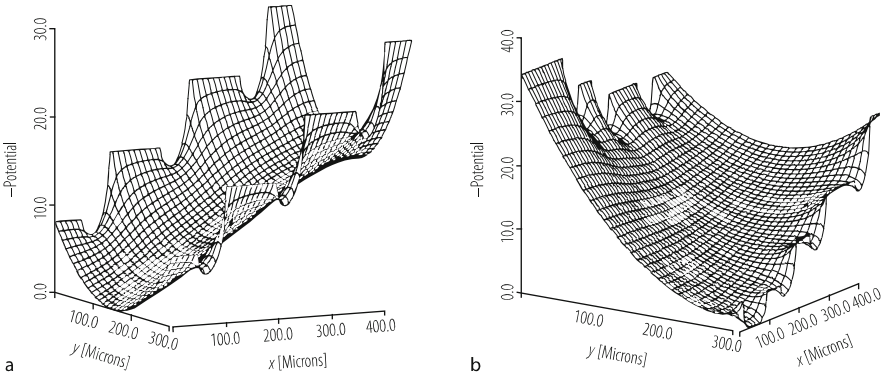


Fig. 5.23 Electron-potential distribution in the linear region (**a**) and close to the anode region where the potential valley is directed towards the surface (**b**)

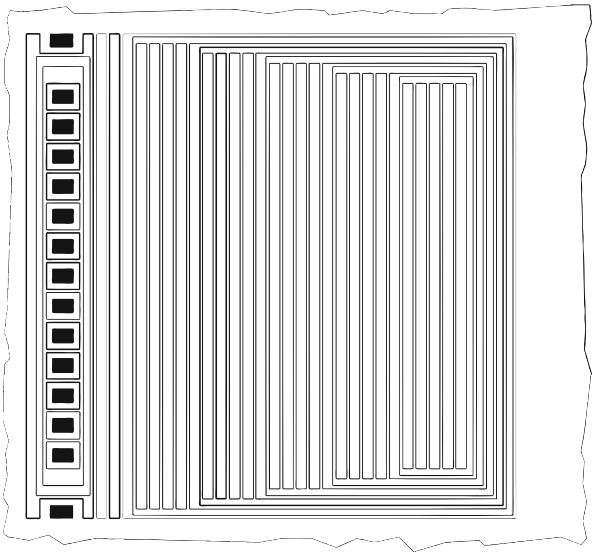


Fig. 5.24 Two-dimensional drift detector with the anode strip divided into pads. The dark pad anodes are embedded in a *p*-doped grid that provides insulation between neighbouring pads

diffusion during the drift time leads to a charge cloud at the anode that is comparable to the spacing of the pads.

For very long drift distances and/or low drift fields, the signal charge will be spread over more than two readout pads. This is an undesirable feature when measuring closely spaced signals. Lateral diffusion can be suppressed by creating deep strip-like p-implanted regions parallel to the nominal drift direction [19]. In this way deviations from the nominal drift direction due to non-uniform doping of the silicon are also avoided.

5.8.2 *Radial and Single Side Structured Drift Devices*

Radial drift devices are in some sense simpler to design than linear devices because the problem of proper termination of the field-shaping strips does not occur. Radial devices are especially interesting for energy measurement. A small point-like anode with extremely small capacitance may be placed into the centre of the device. The small capacitance results in low electronic noise and as a consequence very good energy resolution.

In one special case radial drift to the outside has been realized with a circular anode divided into pads, thus arriving at two-dimensional position measurement in cylindrical coordinates. An interesting feature of such an arrangement is the high position accuracy at small radius in the azimuthal direction. The position in this second coordinate is obtained from the charge distribution measured in the anode pads by projecting it back in the radial direction. A large-area device of this type [20], with a hole in the centre for the passage of the particle beam, has been produced for the CERES particle physics experiment at CERN. The device also uses a method to drain the current generated at the oxide-silicon interface between the field-shaping rings to an *n*-doped drain contact, separated from the signal-collecting anode [21]. In this manner the anode leakage current is reduced and the measurement precision increased.

The S*ilicon Drift Diode* (SDD) [3] combines radial drift with a homogeneous unstructured backside radiation entrance window (Fig. 5.25). Its principal field of application is in (X-ray) spectroscopy where excellent energy resolution is required. A further significant improvement was obtained by integrating a readout transistor into the device (Fig. 5.26). In contrast to the original drift chamber with the electron potential valley located parallel to the wafer surfaces now only one structured surface provides the drift field in the valley which now is at an angle with respect to the wafer surface.

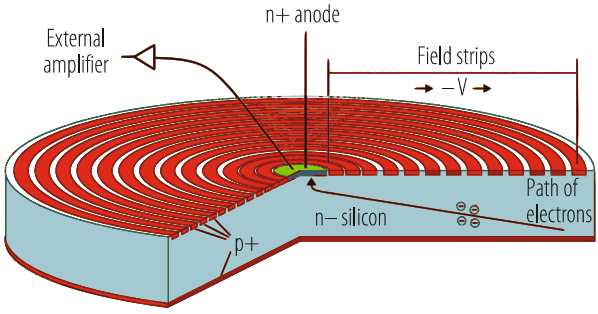


Fig. 5.25 Cylindrical silicon drift detector. The entire silicon wafer is sensitive to radiation. Electrons are guided by an electric field to the small collecting anode in the centre

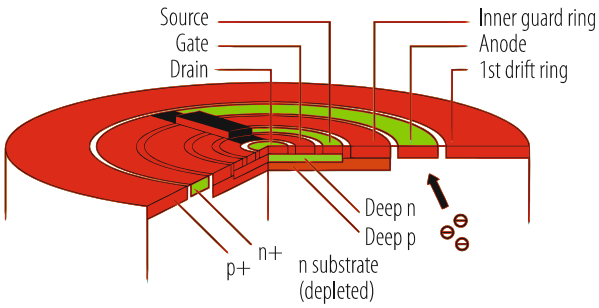


Fig. 5.26 On-chip single sided junction FET coupled to the readout node of a cylindrical silicon drift detector

Having only one surface structured allows using the unstructured surface of the fully depleted device as radiation entrance window. Not having to take other functions into considerations, this radiation entrance window can be made very thin and uniform [22]. The circular geometry with a very small charge-collecting anode in its centre reduces the capacitive load to the amplifier and therefore the noise.

Having the first transistor integrated into the device [23], the capacitance of the detector-amplifier system is minimized by eliminating bond wires between detector and amplifier. In this way stray capacitances between the readout node and ground are avoided, which makes the system faster and less noisy. Further advantages are evident as electrical pickup is significantly reduced and microphony i.e. noise introduced by mechanical vibrations, is excluded. In order to work on the lowly doped and fully depleted substrate, a non-standard "Single Sided Junction Field Effect Transistor" (SSJFET) has been developed [24].

Drift detectors with an integrated transistor are commercially available. They can also be obtained as modules assembled with a Peltier cooler in a gas-tight housing with a thin radiation entrance window (Fig. 5.27). To demonstrate the excellent spectroscopic performance achieved with such devices a spectrum obtained with an ^{55}Fe source and the quantum efficiency are presented in Fig. 5.28 for a cylindrical SDD with a sensitive area of 5 mm^2. The detector temperature, important for the leakage current, was set to −20 °C and the signal shaping time to 1 μs. The $Mn_{K\alpha}$ line at 5.9 keV and the $Mn_{K\beta}$ line at 6.5 keV are clearly separated and their widths are only slightly above the intrinsic Fano limit given by the pair generation process in silicon.

Cylindrical silicon drift diodes with integrated SSJFETs have been manufactured with sensitive areas in the range from 5 mm^2 to 1 cm^2.

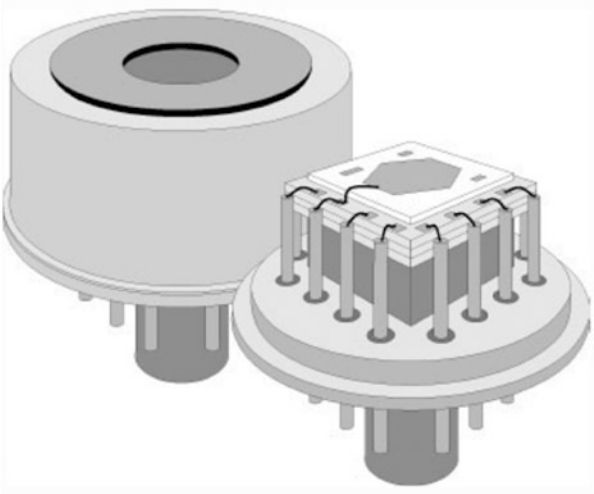


Fig. 5.27 Perspective view of a module consisting of a single-sided structured cylindrical drift detector with integrated SSJFET transistor, cooled by a Peltier element

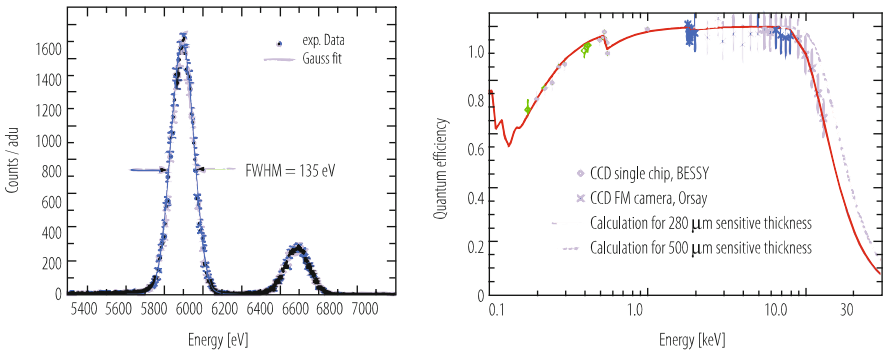


Fig. 5.28 $Mn_{K\alpha}$ – $Mn_{K\beta}$ spectrum (*left*) and quantum efficiency as function of X-ray energy (*right*) of a 5 mm^2 drift diode. The device was operated at −20 °C with a shaping time of 1 μs

5.9 Charge Coupled Devices

Charge coupled devices (CCDs) have for a long time been used as optical sensors, most noticeably as imaging devices in video cameras. Some years ago they also found their application as particle detectors in Particle Physics [25], where specially selected optical CCDs were used. Meanwhile detector systems have been constructed for measuring tracks in electron-positron collisions [26].

p-n CCDs for the special purpose of particle and X-ray detection have been developed [2]. They are based on the principle of side-wards depletion of a double-

diode structure, which is also used in the semiconductor drift chamber. Their first use was in two space-based X-ray telescopes: XMM [27] and ABRIXAS [28].

CCDs are non-equilibrium detectors. Signal charge is stored in potential pockets within a space-charge region, the content of which is then transferred to a collecting readout electrode. In order to retain the thermal non-equilibrium condition, thermally generated charge that also assembles in the potential pockets has to be removed from time to time. Usually this is done during the readout cycle of the device.

While in conventional MOS CCDs minority carriers (electrons in a *p*-type bulk) are collected, the *p*-*n* CCDs are majority carrier (electrons in an *n*-type bulk) devices. The conventional MOS CCDs to be described in the following for didactic purposes store and transfer the charge directly at the semiconductor-insulator interface. These devices are in practice not used anymore and have been replaced by buried-channel CCDs, in which the store-and-transfer region is moved a small distance away from the surface. As a result they are less sensitive to surface radiation damage. In *p*-*n* CCDs, this region is moved a considerable distance into the bulk.

5.9.1 MOS CCDs

The CCD transfer mechanism is explained in Fig. 5.29 that shows a cut along the transfer channel. The top part of the *p*-type bulk is depleted of charge carriers and the potential along the Si-SiO_2 interface is modulated in a periodic fashion with the help of the metal electrodes on top of the SiO_2. Electrons created in the sensitive bulk region assemble in the potential maxima (minima for electrons) at the Si-SiO_2 interface.

The charge can now be moved towards the readout electrode by a periodic change of the voltages ϕ_1, ϕ_2, and ϕ_3, as shown in the figure. First ϕ_2 is increased to the same level as ϕ_1 and the signal charge will spread between ϕ_1 and ϕ_2. If now ϕ_1 is lowered, the signal charge will transfer below the electrodes ϕ_2. If this procedure is followed for ϕ_2 and ϕ_3 and then again for ϕ_3 and ϕ_1, the signal charge is transferred by a complete cell. After several cycles the charge will finally arrive at the anode, where it can be measured.

Placing many of these channels next to each other and separating them by so called channel stops one arrives at a matrix CCD. Channel stops prevent the spreading of signal charge to neighbour channels. They can be realized by doping variations as for example an increased *p*-doping between channels. Usually charge is transferred into one additional charge transfer channel oriented perpendicular to the matrix channel (Fig. 5.30) so that the pixel charge can be shifted towards a single output node.

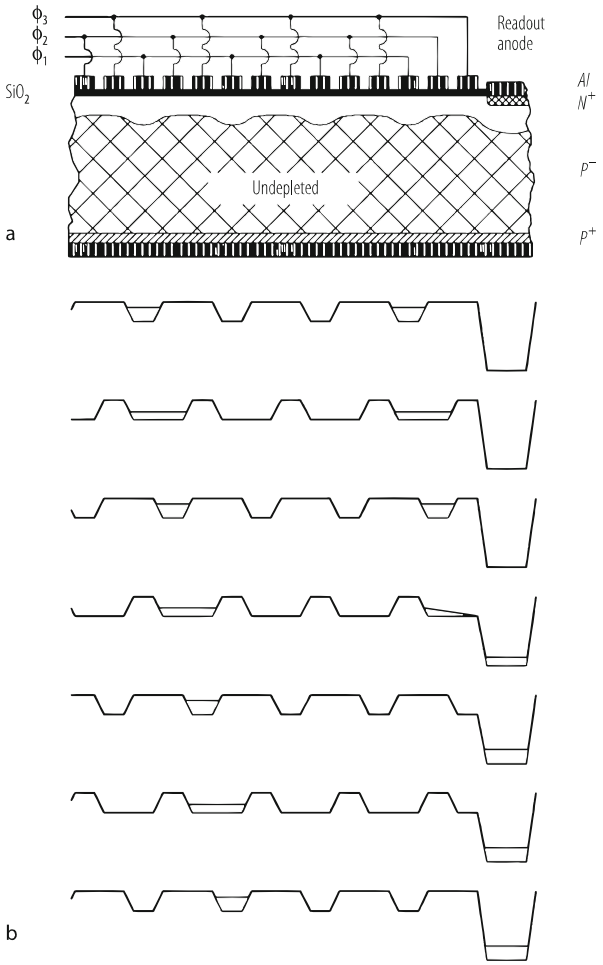


Fig. 5.29 Working principle of a three-phase MOS CCD: layout (**a**); charge-transfer (**b**): Every third gate electrode is connected to the same potential (ϕ_1, ϕ_2, ϕ_3) so that a periodic potential appears below the gates at the Si-SiO_2 interface. Electrons are collected in the maxima of the potential distribution. They can be shifted towards the readout anode by changing the potentials, as shown in (**b**)

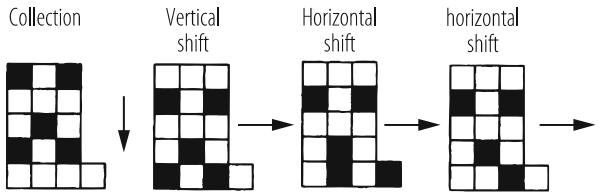


Fig. 5.30 Matrix CCD and the principle of the charge-transfer sequence. Charge is shifted in the vertical direction with all pixels of the matrix in parallel, the lowest row being transferred into a horizontal linear CCD. This horizontal CCD is then read out through a single output node

5.9.2 *Fully Depleted pn-CCDs*

pn-CCDs were originally developed for X-ray imaging in space. A 6×6 cm^2 size device is used as focal imager in one of the three X-ray mirror telescopes at the European XMM/Newton X-ray observatory [29]. From 2000 until the end of the mission in 2018 it has produced high quality X-ray images of the sky [30].

The *pn*-CCD principle, derived from the silicon drift chamber, has already been shown in Fig. 5.5. The layout of the XMM focal plane detector is shown in Fig. 5.31. Twelve 1×3 cm^2 CCDs with 150×150 μm^2 pixel size are monolithically integrated into a single device placed on a 4 inch silicon wafer of 300 μm thickness. Each column of pixels has its own readout channel allowing for fast parallel readout.

Figure 5.32 shows a cross section of a *pn*-CCD along the transfer channel. Here one sees in greater detail the functioning of the device. Contrary to standard MOS-CCDs the registers are formed as *pn*-junctions and the radiation sensitive oxide plays only a minor role. The device is fully depleted with a higher *n*-type doping concentration in the epitaxial layer below the top surface. This leads to a potential distribution shown in the right part of the figure and prevents holes from the p^+-doped registers to be emitted across the wafer towards the backside *p*-doped entrance window. Charge storage and transfer occurs in a depth of approximately 10 μm in contrast to MOS CCDs where this happens at the Si-SiO_2 interface. Fast and efficient charge transfer by drift is therefore possible even for large pixel sizes.

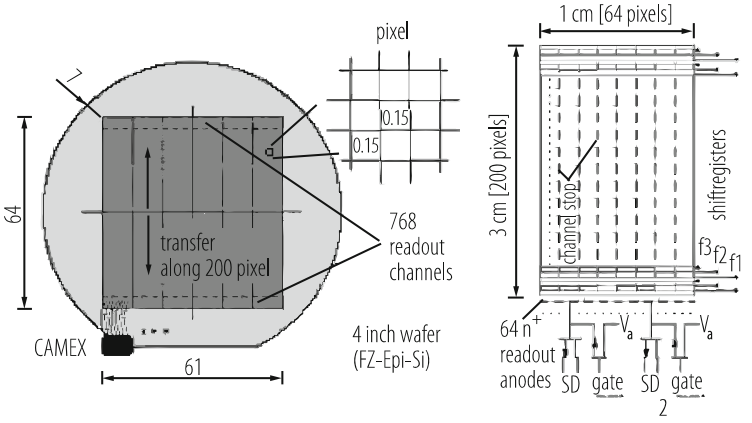


Fig. 5.31 Layout of the XMM *pn*-CCD. 12 logically separate *pn*-CCDs of 1×3 cm^2 area are monolithically fabricated on a 4 in. wafer to a 6×6 cm^2 device with a common backside entrance window. The pixel size is 150×150 μm^2

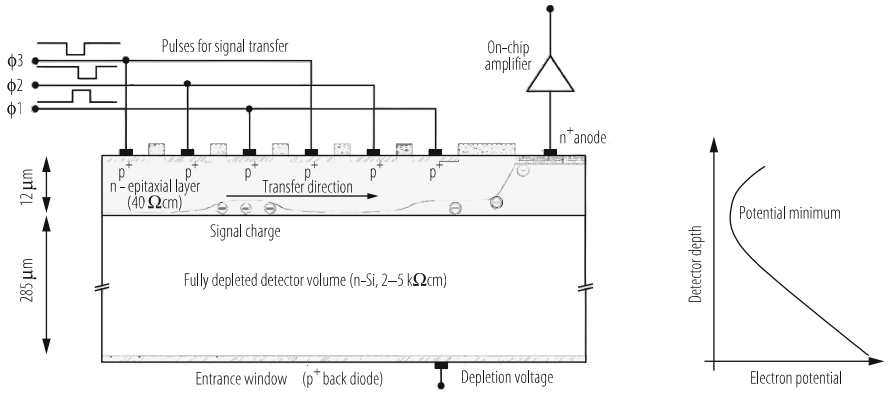


Fig. 5.32 Cross section through the CCD along the transfer channel

5.9.3 CCD Applications

MOS CCDs have a long history in optical imaging. They have been used in camcorders but also in optical astronomy. In particle physics they were first used by the ACCMOR collaboration in the NA11 experiment at CERN where they were successfully employed for heavy flavour decay detection and measurement. They then found their way to collider physics at SLAC and also to X-ray astronomy, where thinning for backside illumination was necessary to achieve sensitivity for low energy X-rays.

Thinning reduces the sensitive volume and therefore the sensitivity at higher X-ray energies. This disadvantage is avoided with *pn*-CCDs that have a typical thickness of 500 µm and, in addition are built with a ultra-thin entrance window so that high quantum efficiency at both low (100 eV) and high (20 keV) X-ray energies is reached. Good radiation tolerance for X-rays is due to two reasons, the absence of sensitive MOS registers and the absorption of X-rays within the bulk before they reach the sensitive charge transfer region (self-shielding). At XMM/Newton *pn*-CCDs have been operating in space for 18 years without noticeable performance degradation.

Compared to MOS CCDs the readout speed is significantly increased due to the larger pixel size, the higher charge transfer speed and parallel column readout. Very large pixel sizes cannot be realized in MOS CCDs that transfer charges very close to the Si-SiO_2 interface.

Use in a further X-ray mission is in preparation: eROSITA (extended ROentgen Survey with an Imaging Telescope Array). Here the CCD is split into an image collecting area and a frame store area. After collection, the complete image is transferred very fast into the frame store area from where it is read with moderate speed row by row while at the same time the next image is collected. The typical image frame readout takes 1 ms, while for MOS CCDs it is in the range of 1 s.

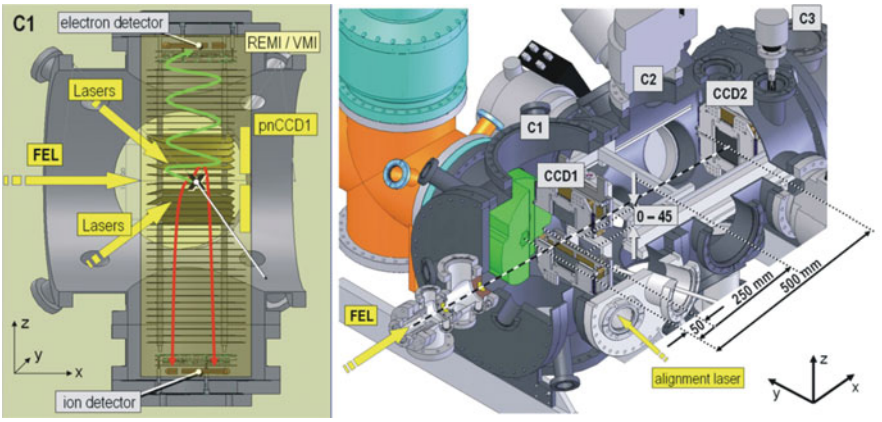


Fig. 5.33 Schematic section through the CAMP detector. The reaction electron and ion detectors with the first CCD sensor plane are depicted on the left hand side. The *pn*-CCD detectors shown in perspective view on the right can detect all photons emerging from the target. In addition, the design allows feeding in other lasers for alignment or pump-probe purposes, as well as for mounting other high-resolution, small-solid-angle electron TOF or crystal spectrometers. The pnCCD1 can be moved in all three directions with a maximum distance of 25 cm along the beam trajectory

Although *pn*-CCDs have been developed for X-ray astronomy they are also visible-light detectors. One application is in adaptive optics that corrects in real time mirror geometries of optical telescopes in order to compensate for atmospheric turbulences at frequencies of approximately 1 kHz.

pn-CCDs are also used in experiments at accelerator-based light sources in particular at X-ray Free Electron Lasers (e.g. FLASH and the European XFEL at Hamburg and LCLS at SLAC). The Center of Free Electron Science (CFEL) in Hamburg has designed the CFEL-ASG Multi Purpose (CAMP) chamber (Fig. 5.33) [31], which combines electron and ion momentum imaging spectrometers with large area, broadband (50 eV to 25 keV), high dynamic range, single photon counting and imaging X-ray detectors based on *pn*-CCDs. The excellent low energy response of *pn*-CCDs has been demonstrated by measuring the response to 90 eV photons at FLASH (Fig. 5.34).

5.10 Active Pixel Detectors

The CCDs discussed in the previous chapter collect charges in pixels during their charge collection period and transport them during the transfer period pixel by pixel to a readout node. Charges produced during the transfer cycle will also be read but the assigned position will be wrong. In active pixel detectors each pixel has its own readout channel and the charge will be assigned to the pixel where it was generated. There are four types of active pixel detectors:

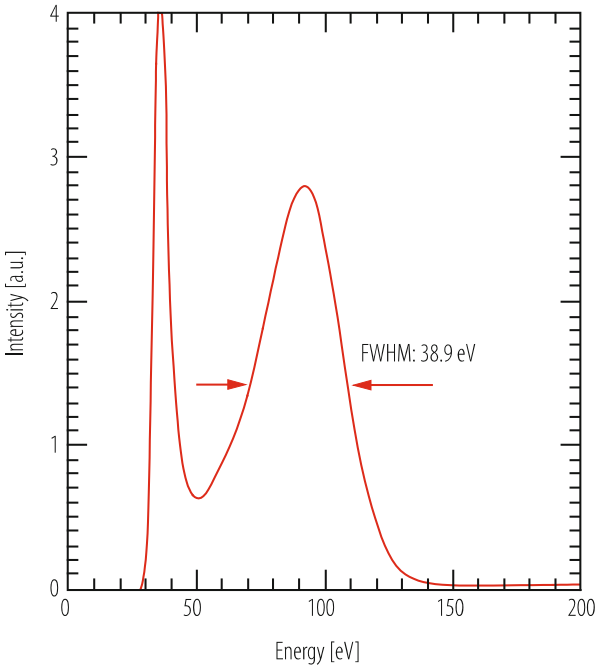


Fig. 5.34 Energy resolution measured at FLASH with 90 eV photons. Every photon generates approximately 25 electron-hole pairs, which are detected with a read-out noise of 2.5 electrons (rms). The measured FWHM energy resolution is only 38.9 eV

(a) Hybrid pixel detectors are diode arrays bonded to an electronics chip produced on a separate wafer so that each pixel has its own readout channel.
(b) MAPS (Monolithic Active Pixel Sensors) are pixel arrays with readout for every pixel directly integrated on the same chip.
(c) DEPFET pixel detectors are two dimensional arrays of DEPFETs with parallel charge collection in the DEPFETs and serial delayed readout of the charges stored in the internal gates.
(d) DEPFET Macro Pixel detectors, pixel detectors with large cell size combine DEPFETs with drift detectors.

All these detector types exist in many variations. Hybrid pixel detectors and MAPS allow parallel data processing and can perform complex tasks thanks to the miniaturized VLSI electronics. This however has a price in power consumption. DEPFET pixel detectors so far are built in a technology of moderately large feature size. Thus complex data processing is not foreseen. Its advantages are sensitivity over the whole bulk, high energy resolution and very low power consumption.

5.10.1 Hybrid Pixel Detectors

Hybrid pixel detectors are used at the Large Hadron Collider (LHC) as the tracking detectors closest to the beam, where the track densities is highest and the radiation exposure most severe. They also became a standard detector for X-ray imaging, in particular at accelerator driven X-ray sources. In their simplest form they consist of a detector wafer with a two dimensional diode array and separate electronics wafers as shown in Fig. 5.35. Every diode is individually connected by bump bonding to its own readout channel. Other connection techniques, including capacitive coupling, have been demonstrated. As readout and sensor are separate, the sensor material can be freely chosen, e.g. a high-Z sensor for the detection of high-energy X-rays.

The main challenge in such a device lies in the electronics that has to provide several functions as for example low noise charge readout and high dynamic range, and—depending on the application—data storage, zero suppression and transmission to the external electronics in analogue or digital form. These functions have to be implemented on an area of the pixel size. Frequently very high speed operation at low power is required as is the case for example in the LHC at CERN. Reaching these goals has been possible by profiting from the dramatic industrial progress in submicron electronics and adapting it to the specific needs. The use of submicron electronics that uses very thin gate oxides has also alleviated the problems with respect to radiation damage.

The typical pixel dimension for the hybrid pixel sensors presently operating at the CERN LHC are of order $100 \times 100\ \mu m^2$. The modules of the ATLAS vertex

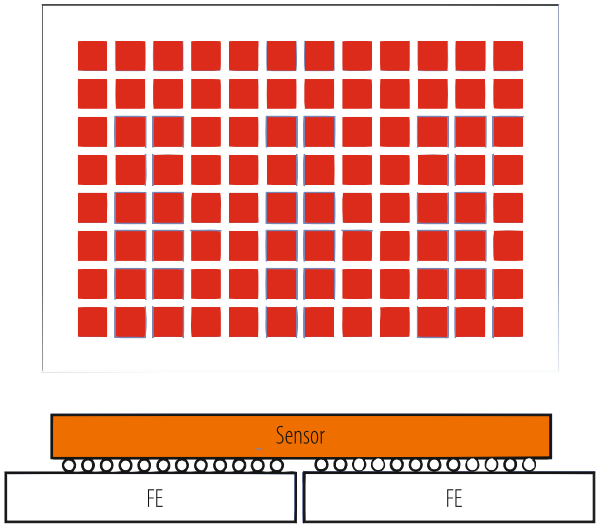


Fig. 5.35 Concept of a Hybrid Pixel Detector consisting of a diode array "flip chip" bonded to several readout chips

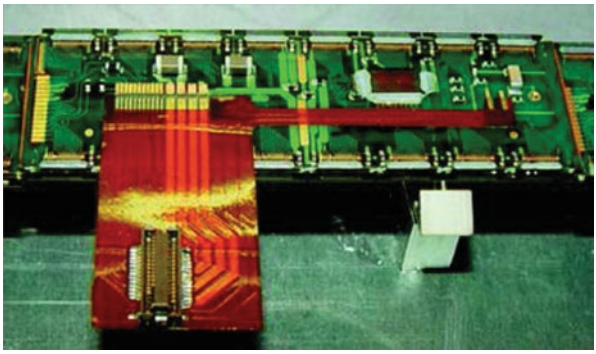


Fig. 5.36 Photo of an ATLAS pixel detector module

detector, shown in Fig. 5.36, have a pixel size of 50 μm × 250 (400) μm, the ones of CMS 100 μm × 150 μm For the High-Luminosity LHC hybrid pixel detectors with pixel sizes of 50 μm × 50 μm and 25 μm × 100 μm are under development.

The hybrid pixel detectors used for X-ray science face somewhat different challenges and follow different concepts. AGIPD (Adaptive Gain Integrating Pixel Detector) [32], which operates at the European XFEL at Hamburg, where X-rays are delivered in pulse-trains with 220 ns distance between pulses, is designed to detect single and up to 10^4 photons with energies in the range 5–15 keV per pulse in pixels of 200 μm × 200 μm, and store 350 frames to be read out in between the pulse trains. This is achieved by signal-driven switching into four gain ranges. In addition, the 500 μm thick pixel sensor is designed for a breakdown voltage above 900 V for ionizing doses up to 1 GGy. There are many applications in X-ray science, where the recording of individual frames is not required, but the number of hits above a given threshold or in a given energy interval are counted for every pixel or the integrated charge for a given time interval recorded. As the electronics takes significantly less space than required for recording and storing individual frames, pixel sizes as small as 55 μm × 55 μm have been achieved. Outstanding examples for such detectors are PILATUS [33] developed at PSI, and the MEDIPIX series [34], developed by a collaboration centred at CERN.

5.10.2 Monolithic Active Pixel Sensors (MAPS)

This name is used for pixel sensors produced with integrated circuit technology on a single wafer using part of the substrate as detector material. One advantage of MAPS is the significantly easier fabrication of detector modules resulting in a significant cost reduction; another is that MAPS can be produced in CMOS Fabs, which includes a fast turn-around time for the development. However, MAPS are very complex devices and achieving all the requirements of the experiments at high-luminosity, including their radiation performance remains a challenge.

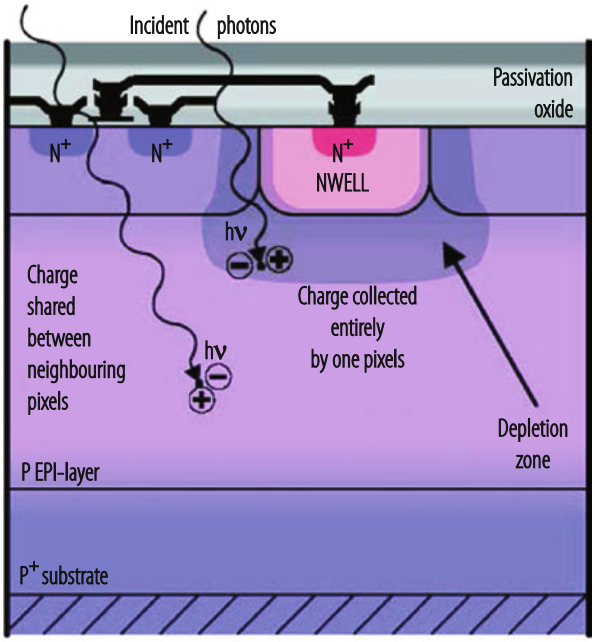


Fig. 5.37 Cross section through a pixel of a MAPS fabricated on CMOS technology but using only NMOS transistors

A first successful demonstration of MAPS operating in an experiment is the EUDET beam telescope [35], with MAPS using only *n*-channel transistors out of an original CMOS technology. Figure 5.37 shows the cross section through a MAPS pixel cell. The *n*-well is used as collecting electrode and all transistors are placed within the *p*-wells. A small volume next to the *n*-well is depleted of charge carriers. In this region signal electrons are collected by drift, but, the major part of the sensitive volume—the *p*-epitaxial layer—is field-free. Thus most of the charge is collected by diffusion, which is intrinsically slow and leads to a large spread of charge into neighbouring cells. There are good reasons why *p*-type transistors are avoided. They would have to be placed into an *n*-well. If this well were separated from the charge collecting electrode it—depending on the *n*-well potentials—would collect signal electrons in competition to the signal electrode or might even inject electrons into the bulk. If it were put into the same well as the collecting electrode it would induce charge directly into the input of the pixel.

For photon detection—as shown in the figure—in addition the material on the top as for example the conducting leads as well as the thick insensitive well zones will absorb part of the incident radiation.

The pixel circuitry (Fig. 5.38) is rather simple. It consists of an NMOS input transistor, a reset transistor and an output select switch. Signal charge is stored at the

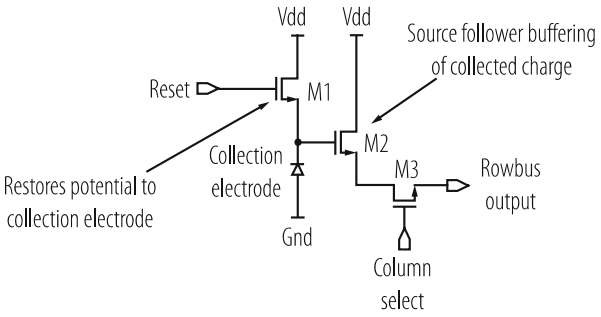


Fig. 5.38 Pixel circuitry of MAPS based on CMOS technology but using only three NMOS transistors. The collecting electrode is directly connected to the gate of a source follower (M2) whose load is common to all pixels of a column and activated by the column select switch. The input node is reset with the reset transistor M1

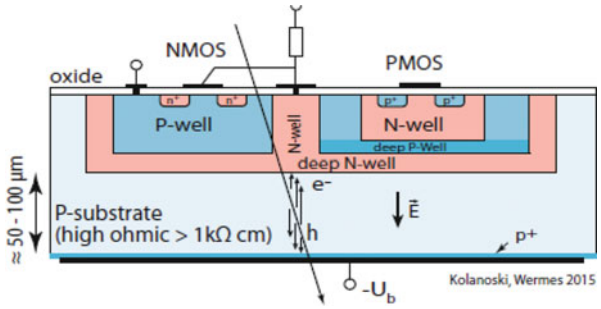


Fig. 5.39 DMAPS with large collection electrodes (figure from Wermes-Kolanoski)

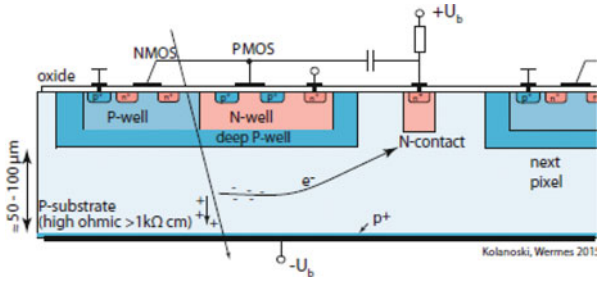


Fig. 5.40 DMAPS with small collection electrodes

input node, read out sequentially and cleared afterwards. MAPS using both CMOS types have also been developed [36].

To overcome the problem of slow charge collection by diffusion, which also makes the sensor sensitive to bulk radiation damage, DMAPS (Depleted CMOS Active Pixel Sensors), are being developed [37]. They are fabricated on substrates with resistivity between 100 Ω·cm and a few kΩ·cm and operated with depletion depths of typically 50–200 µm. As shown in Figs. 5.39 and 5.40, two approaches

are followed: Large Collection Electrode (a) and Small Collection Electrode (b). Design (a) has the advantage of a more uniform electric field resulting in shorter drift distances, and thus a good radiation tolerance is expected. Its disadvantage is the large capacitance of about 100 fF per pixel and an additional well-to-well capacitance of similar value, which results in increased noise, reduced speed, higher power consumption and possibly cross-talk between sensor and digital electronics. Design (b) has a small electrode adjacent to the well in which the electronics is embedded. This has the advantage of a small capacitance of about a few fF and thus improved noise and speed at low power. However, the electric field in the sensor is not uniform with low field regions. This makes them more sensitive to radiation damage. DMAPS of both types have been fabricated by different foundries in 150 nm, 180 nm and 350 nm technologies. They show impressive results even after irradiation with hadrons to fluences exceeding a few 10^{15} cm^{-2}.

5.10.3 DEPFET Active Pixel Sensors

The Depleted Field Effect Transistor structure shown in Fig. 5.6 is a natural building element for a pixel detector. It acts simultaneously as detector and as amplifier. A variety of DEPFET designs can be constructed. Figure 5.41 shows two examples, one with cylindrical, the other with linear geometry.

Arranging many of these devices in a matrix and connecting them in such a way that selected DEPFETs can be turned on, one arrives at a pixel detector with charge

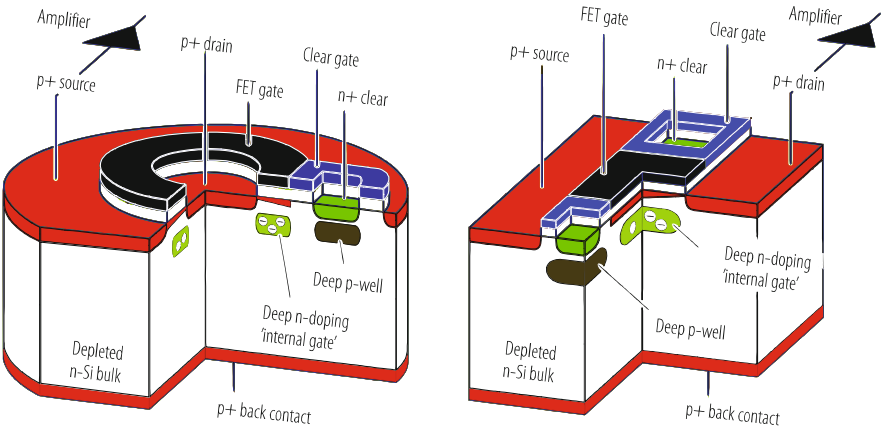


Fig. 5.41 Schematic drawings of MOS-type DEPFETs with circular (*left*) and linear (*right*) geometry. The signal charge is collected in a potential well ("internal gate") below the FET gate, thereby increasing the conductivity of and thus the current in the transistor channel. The collected charges can be drained towards the clear contact by applying voltage pulses to the clear contact and/or the clear gate

storing capability. Before turning to the matrix arrangement the main properties of the DEPFETs are summarised:

- Combined function of sensor and amplifier;
- Full sensitivity the over complete wafer, low capacitance and low noise, non-destructive repeated readout, complete clearing of the signal charge and thus no reset noise.
- Continuous (real time) and integrating (charge storage) operating modes can be chosen.

The signal can be read out either at the source as indicated in the left figure or at the drain as shown in the linear example. With source readout one compensates the increase of channel conduction due to the charge in the internal gate by a reduction of the external gate-source voltage, seen as voltage change of the source. In the drain readout the source potential is kept constant and the drain-current change can be directly observed. An important property in pixel detector applications is the fact that the signal charge collection occurs not only for current carrying DEPFETs but also for those which have been turned off with the help of the external FET gate.

DEPFET pixel sensors have been developed at the MPI Semiconductor Laboratory in Munich for several purposes, as focal sensors of the proposed European X-ray observatory XEUS [38] and as vertex detector for the BELLE-II experiment at KEK in Japan and the proposed International Linear Collider ILC. In XEUS the combined functions of imaging and spectroscopy are of importance, for the vertex detectors the measurement of position of charged tracks is of prime interest. This however has to be done with very high precision (few μm) and at high readout speed. The position measurement requirement in XEUS is not as stringent; it is matched to the expected quality of X-ray imaging. However, highest emphasis is given to spectroscopic quality and quantum efficiency and data readout speed is still large.

As a consequence of these and further requirements circular geometries have been chosen for XEUS and linear ones for the vertex detectors (see Fig. 5.41). The excellent spectroscopic capabilities of DEPFETs can be appreciated from the ^{55}Fe source spectrum taken with a single circular pixel cell (Fig. 5.42).

The DEPFET with its capability of creating, storing and amplifying signal charge is an ideal building block for a pixel detector. A large number of DEPFETs can be arranged in a matrix in such a way as to power selected DEPFETs for reading and clearing the collected signal charge. Figure 5.43 shows a rectangular arrangement of DEPFETs. Their drains are connected column wise while gates and clear electrodes are connected row wise. Each row has its individual readout channel. A row at a time is turned on with the help of the gate voltage while all other DEPFETs have zero current. Charge collection does not require a current within the DEPFET.

Readout can be performed in double correlated mode: Turning on the current with a negative voltage on the gate is followed by a first reading of the current, a clearing of the signal charge in the internal gate with a positive pulse at the clear contact and a second current reading before the current is turned down again and reading is switched to the next row. The difference of first and second current

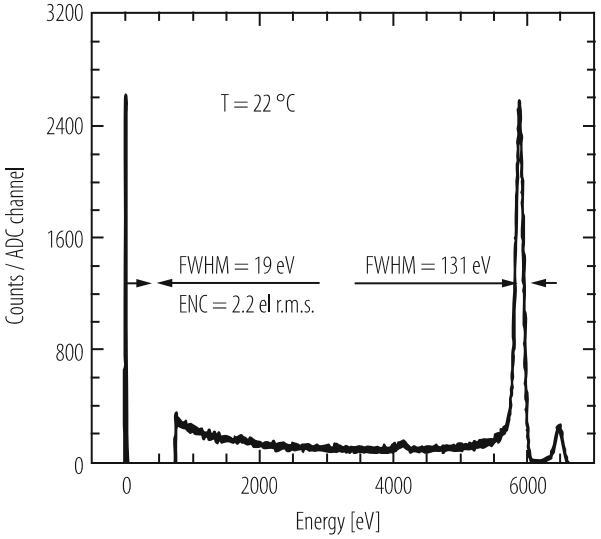


Fig. 5.42 ^{55}Fe spectrum measured with a single circular (XEUS-type) DEPFET. A spectral resolution of 131 eV has been obtained with room temperature operation and 6 μs Gaussian shaping. The separately measured noise peak has a FWHM of 19 eV corresponding to an electronic noise of 2.2 electrons r.m.s

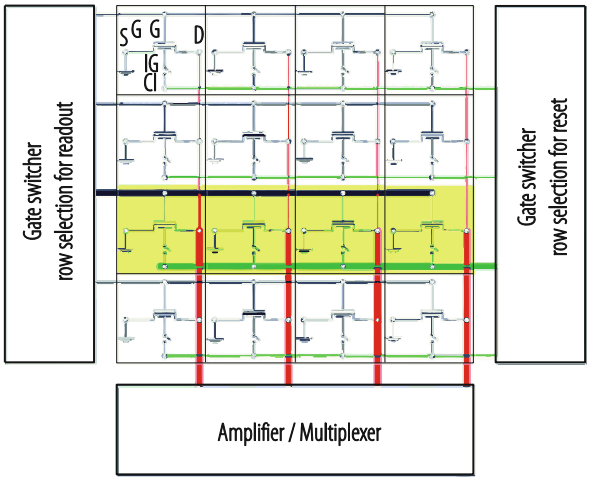


Fig. 5.43 Circuit diagram of a DEPFET pixel detector with parallel row-wise readout of the drain current

reading is a measure for the signal charge in the pixel cell. Alternatively to the procedure described above, sources may be connected column wise and source voltages measured instead of drain currents. Figure 5.44 shows the spectroscopic quality reached with a 64 × 64 DEPFET matrix of 50 × 50 μm^2 pixels.

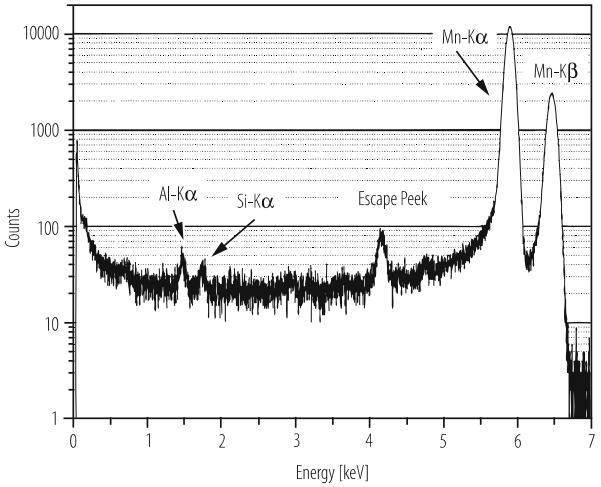


Fig. 5.44 ^{55}Fe spectrum measured at −28 °C with a 64 × 64 cell DEPFET pixel matrix with 50 μm pixel size

Pixel sensors with large pixels can be constructed by combining DEPFET structure and drift chamber principle. Large pixel may be preferred in order to increase the readout speed and reduce the number of readout channels and power consumption. It is advisable to match the pixel size to the properties of the rest of the system. Over-sampling may increase the electronic noise lead to a worse performance.

Macro Pixel DEPFET Sensors

Figure 5.45 shows the principle with a cut and a top view of a cell. The circular DEPFET structure is located in the centre of a cylindrical drift detector. Electrons created anywhere in the fully depleted bulk are driven by the suitably shaped drift field towards the internal gate below the transistor channel. For this device a new type of DEPFET has been invented that allows clearing of the signal charge with substantially lower voltage by putting the clear electrode inside the drain region located in the centre. The drain region does not consist of a highly doped *p* region but is formed by an inversion layer that is controlled by a gate voltage and automatically connected to the small drain contact. Putting a sufficiently high positive voltage on this gate, the drain assumes the role of the clear electrode, which is automatically connected to the *n*-doped clear contact.

Single pixel cells and a 4 × 4 1 mm^2 pixel matrix (Fig. 5.46) have been tested successfully. Figure 5.47 shows an ^{55}Fe spectrum taken at room temperature. Here one notices a somewhat worse spectroscopic resolution than with the small-pixel devices. This is due to the leakage current which now is collected from a volume which is larger by a factor 400. The leakage current can be suppressed by lowering the operating temperature.

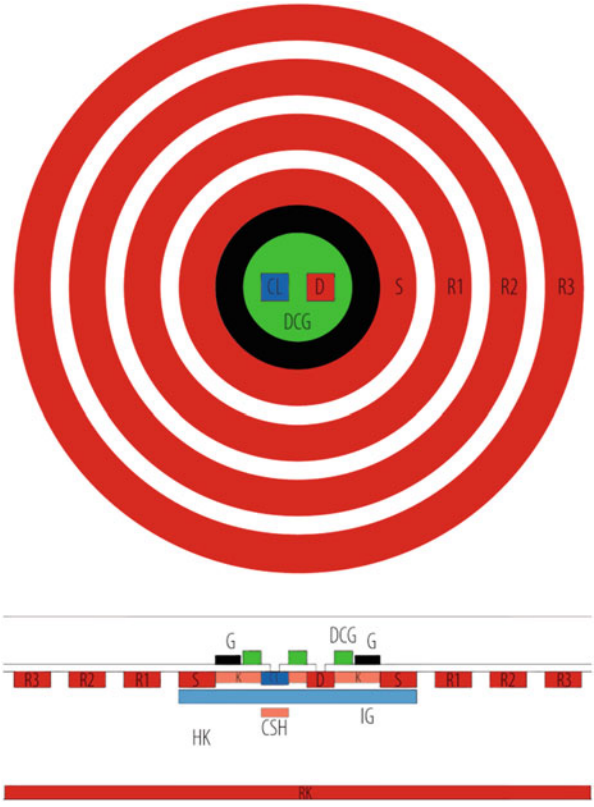


Fig. 5.45 Principle of a macro-pixel cell: A DEPFET located at the centre of a drift detector serves as storage and readout device

New DEPFET Developments

The DEPFET concept allows a variety of further functionalities that have partially been proven experimentally but not yet implemented into a large area pixel detector:

(a) As signal charge is not destroyed by the readout process this charge can be read repeatedly and the measurement precision improves with the square root of the number of measurements. This has been verified with a pair of neighbouring DEPFET transistors arranged in such a way as to allow the transfer of signal charge from one internal gate to the other and in reverse direction. A measurement precision of 0.25 electrons has been achieved independently of the amount of signal charge [39].

(b) Gatable DEPFETs [40] are developed for applications in High Time Resolution Astronomy (HTRA) and Adaptive Optics. They collect signals in preselected time intervals only, whereas the charge generated outside of these gate periods are drained towards a clear electrode.

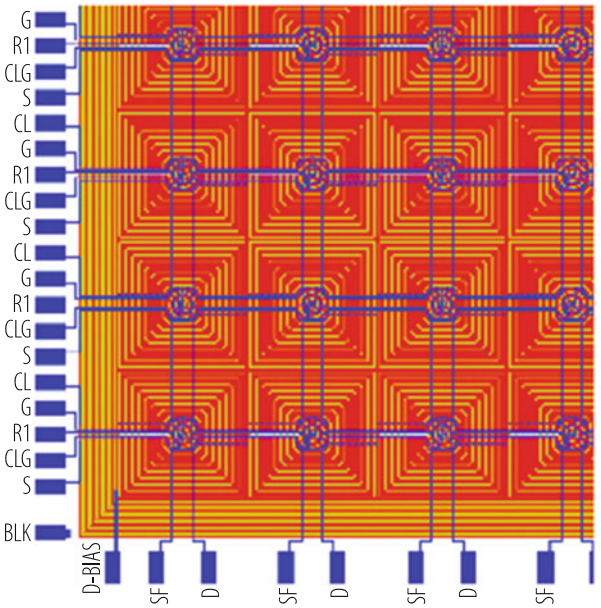


Fig. 5.46 Layout of a macro pixel matrix

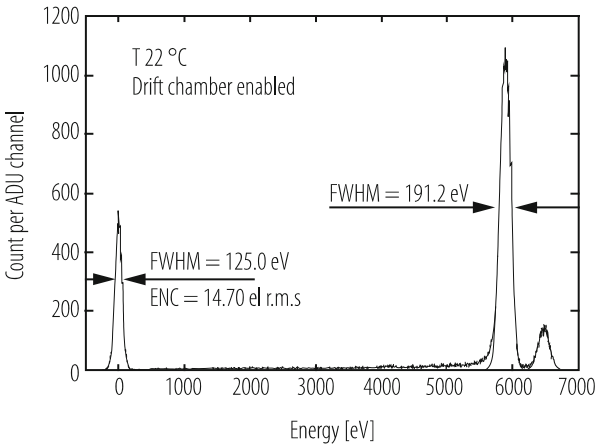


Fig. 5.47 ^{55}Fe spectrum measured at room temperature in a 1×1 mm^2 pixel of an 8×8 macro pixel matrix with 6 μs shaping. The increase of the noise compared to single DEPFET cells is due to the leakage current in the large sensitive volume of $1 \times 1 \times 0.45$ mm^3, which can be reduced by cooling

(c) Nonlinear DEPFETs [41] developed for applications at the European X-ray Free Electron Laser (EuXFEL) at Hamburg. Their non-linear characteristics and high-speed capability combines simultaneously single X-ray-photon sensitivity and very high dynamic range at the 5 MHz EuXFEL repetition rate.

DEPFET Pixel Detector Applications

In the last years DEPFET pixel detectors have been developed at the MPI Semiconductor Laboratory for the following projects:

Bepi Colombo, a mission for observing mercury [42], XEUS/IXO a space based X-ray observatory that will succeed the XMM/Newton and vertex detectors for the International Linear Collider (ILC) and the BELLE-II experiment at the KEK e^+e^- collider.

As an example for the application in X-ray detection Fig. 5.48 shows spectra at high readout rates taken with a Bepi Colombo prototype macro pixel detector. In the final detectors (Fig. 5.49) the pixel size is reduced to $300 \times 300\ \mu m^2$. An X-ray image obtained by illumination through a mask (Fig. 5.50) demonstrates functioning of the full detector.

The ILC and BELLE vertex detectors [44, 45] require fast readout (10 μs frame time), excellent spatial resolution (5 μm) and minimal material thickness to

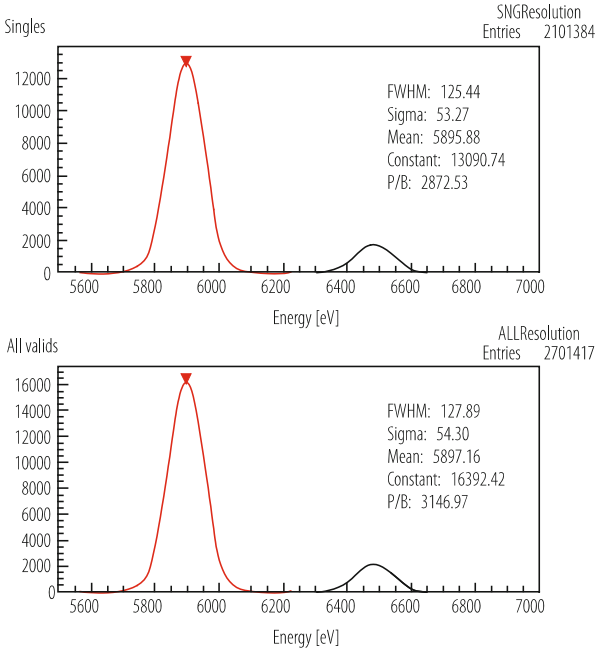


Fig. 5.48 Spectroscopic resolution of Bepi Colombo macro-pixel detectors with 64 × 64 pixels of 500 × 500 μm^2 size on a 500 μm fully depleted substrate with ultra-thin backside radiation entrance window. The top figure is restricted to photons contained in single pixels. while in the lower part signals split between neighbour pixels are included. Readout was with the ASTEROID pixel chip [43] that averages the DEPFET signals over an "integration time" once before and once after clearing and takes their difference as a measure for the deposited charge. The measured width of 125 eV FWHM with 0.9 μs integration time corresponds to an electronic noise of 4 electrons r.m.s. Reducing the integration time from 0.9 to 0.25 μs increases the width to 163 eV FWHM corresponding to 13 electron charges r.m.s

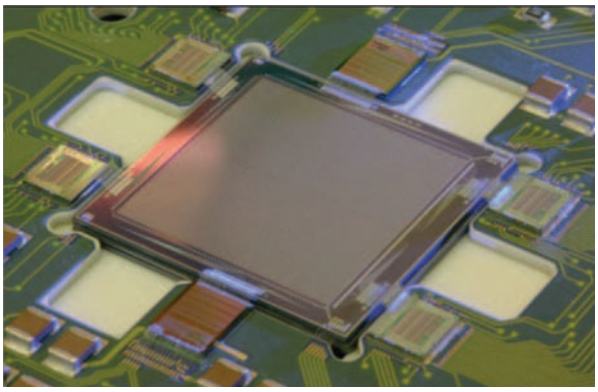


Fig. 5.49 Photo of an assembled macro-pixel detector with two 64 channel ASTEROID readout chips on top and bottom and four steering chips

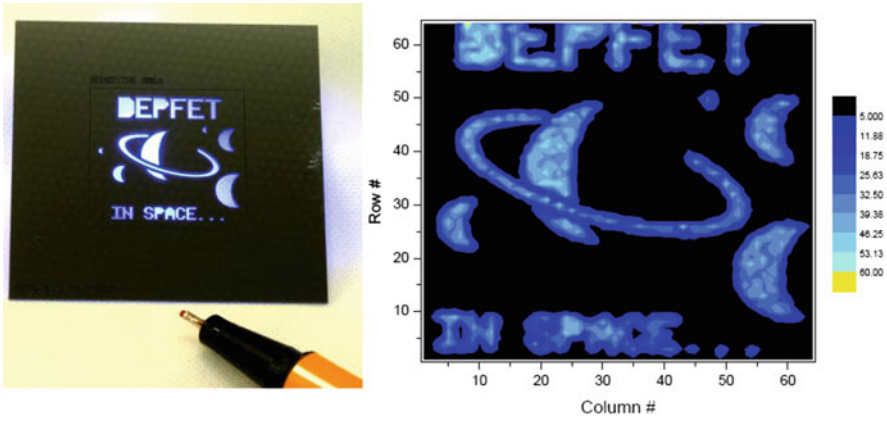


Fig. 5.50 X-ray image (right) obtained with the mask shown on the left

minimize the scattering of charged particles. Consequently the pixel size has been chosen as $25 \times 25\ \mu m^2$ for ILC and $50 \times 75\ \mu m^2$ for BELLE-II. A new method for wafer thinning based on wafer bonding technique has been developed in order to produce thin (50 μm) self-supporting all silicon modules [46].

5.11 Detectors with Intrinsic Amplification

Contrary to gas detectors, semiconductor detectors usually provide only the primary ionization as signal charge. This mode of operation is possible because of the low energy needed for producing an electron-hole pair (3.6 eV in silicon, whereas the ionization energy for gases is about 30 eV) and the availability of low noise

electronics. The measurement of the primary ionization without gain avoids any effect of gain variation or amplification noise, and thus leads to stable operation in spectroscopic measurements. However, high speed and very low noise requirements, detection of single photons, compensation for charge losses due to radiation damage or timing accuracies of the order of tens of picoseconds, make an intrinsic amplification of the detectors desirable.

A rather old and well known device is the avalanche diode, with several different operating modes. In the last two decades arrays of avalanche diodes operated in the Geiger mode (SiPMs—Silicon Photo Multipliers) have become photo-detectors of choice for many applications, and more recently tracking detectors with gain (LGAD—Low Gain Avalanche Detectors) are developed with the aim to combine precision position with precision timing in the harsh radiation environment of the high-luminosity LHC at CERN.

5.11.1 Avalanche Diode

An avalanche diode has a region with a field of sufficient strength to cause charge multiplication. An example of such a device is shown in Fig. 5.51. The base material is low doped p-type silicon. The junction, consisting of a thin highly doped n-type layer on top of a moderately doped p-layer, may also be used as entrance window for radiation, especially when the bulk material is only partially depleted.

An enlarged view of the central top region of Fig. 5.51, in which multiplication takes place, is shown in Fig. 5.52. Also shown are charge density, electric field and potential for the idealized assumption of uniform doping in the n^+-, p- and p^--regions ignoring diffusion. The middle p region is fully depleted and the space-charge region extends into the thin n^+ top region and the low doped p^--bulk. The maximum of the electric field is at the n^+p junction.

Electrons produced below the n^+p junction (and holes produced above the junction) will pass the high field region of the junction when drifting in the electric

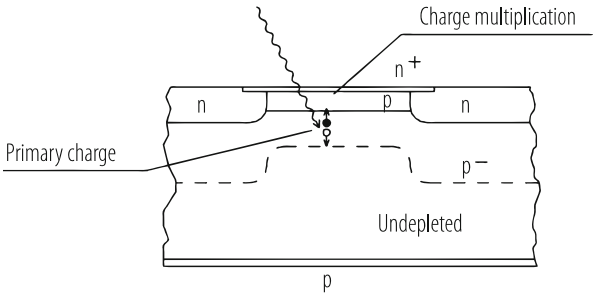


Fig. 5.51 Avalanche diode built on p-type silicon with a high-field region right below the top surface

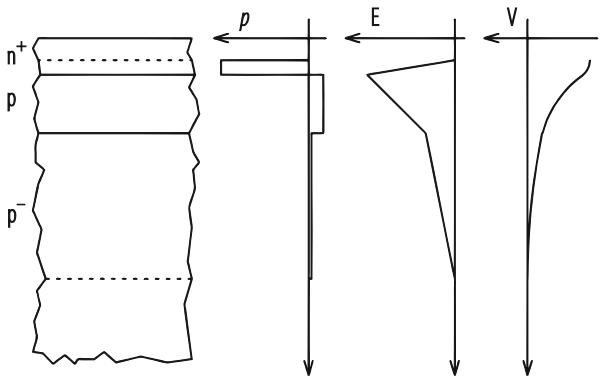


Fig. 5.52 Amplification region of the avalanche diode shown in Fig. 5.51. Also shown are charge density ρ, electric field E, and potential V

field towards the collecting electrode on top (on bottom). If the electric field is strong enough to accelerate electrons (or holes) between collisions with the lattice imperfections so that the kinetic energy is sufficient to create another electron-hole pair, the charge produced by the primary ionization is amplified.

One important aspect to be considered in designing or operating avalanche diodes is the different behaviour of electrons and holes with respect to charge multiplication. In silicon, the onset of charge amplification for holes occurs at higher electric fields than for electrons. The situation is opposite in germanium, while in GaAs the difference between electrons and holes is comparatively small.

Therefore several working regimes exist that vary depending on the strength and extension of the high electric field region. In the case of silicon one finds: (a) At low electric field, no secondary electron-hole pairs are generated. The device has the characteristics of a simple diode. (b) At higher electric field only electrons generate secondary electron-hole pairs. The amplified signal will be proportional to the primary ionization signal, with some statistical fluctuation from the multiplication process added to the fluctuation in the primary ionization process. (c) At even higher field, holes will also start to generate secondary electron-hole pairs. Secondary electrons generated by holes will again pass through (part of) the amplification region, thereby possibly generating other (tertiary) electron-hole pairs. This avalanche process will continue until it is either stopped by a statistical fluctuation in the multiplication process or by a sufficiently large drop of the externally supplied voltage. This drop may be due to the increased current passing through a bias resistor or an external enforcement by, for example, a feedback circuit. The generation of a large number of free charge carriers in the multiplication region also reduces the electric field strength and therefore decreases charge multiplication in later stages of the avalanche generation. In this operation mode the output signal is no more proportional to the primary charge; however, single photon detection becomes possible.

5.11.2 Low Intensity Light Detection

An optical photon in its primary interaction will create a single electron-hole pair, a charge too small to be detected by standard electronics. However, intrinsic amplification in an avalanche process makes single photon detection possible. The avalanche diode of Fig. 5.51 is such a device. Operation in proportional mode will result in an output signal proportional to the number of (optical) photons, with some statistical fluctuations of the avalanche process added and additional contributions from the non-uniformity of the electric field in the avalanche region. Operation in limited Geiger mode will result in a signal independent of the number of incident photons. The charge signal will be approximately given by the product of the diode capacitance times the difference of the applied voltage and the voltage at which the avalanche process stops.

As the charge multiplication probability is a strong function of the electric field strength, high uniformity over the active area is required and high field regions at the edge of the device have to be avoided by proper design. Edge breakdown is avoided in Fig. 5.51 by the less strongly doped n region at the rim. This leads to a space-charge region extending deeper into the bulk and to a reduction of the maximum field.

If the structure of Fig. 5.51 is to be operated in proportional mode (with only electrons multiplying), primary charge produced by radiation entering from the top has to be generated below the high field multiplication region in order to be properly amplified. Therefore for blue light, with its submicron penetration depth, the efficiency is low for this design.

In choosing the width of the depleted region, one has to consider several partially conflicting requirements. Based on noise considerations, this region should be large in order to reduce the capacitive load to the amplifier. The same is required for the detection of deeply penetrating radiation such as X-rays or energetic charged particles. One may even extend the depleted region all the way to the bottom surface. Then the backside p-doped surface can also be used as a radiation entrance window. This can be an advantage for low penetrating radiation such as optical photons, since such an entrance window can be made thin. The disadvantage of a large depleted region is the large volume for thermal generation of electron-hole pairs, the electrons being capable of initializing the avalanche process and, depending on the application, a not wanted sensitivity to deeply penetrating radiation.

The electric field configuration in the avalanche region is shown in an idealized way in Fig. 5.52, assuming abrupt doping changes. Such a distribution is not only unrealistic but also far from optimal for proportional operation: Breakdown should be avoided as much as possible which can be achieved by an extended amplification region and lower hole-to-electron multiplication ratios, as is the case for lower fields. Such a design can be realised by suitably doping the avalanche region.

5.11.3 Solid-State Photo Multipliers: SiPMs

In the last decade a new type of avalanche photon detector has reached maturity and is now commercially available, the Solid State Photo Multiplier, also referred to as SiPM (Silicon Photo Multiplier), G-APD (Geiger Mode Avalanche Photo Diode) or MPPC (Multi Pixel Photon Counter) [47]. It consists of two dimensional arrays of 100–10,000 single photon avalanche diodes (SPADs), called pixels, with typical dimension between $(10\ \mu m)^2$ and $(100\ \mu m)^2$. The pixels are operated in limiting Geiger mode and every pixel gives approximately the same signal, independent of the number of photons which have produced simultaneously electron-hole pairs in the amplification region of the pixel. The sum of the pixel signals is equal to the number of pixels with Geiger discharges, from which the number of incident photons can be determined. As the output charge for a single Geiger discharge is typically larger than 10^5 elementary charges, 0, 1, 2, and more Geiger discharges can be easily distinguished, enabling the detection of single optical photons with high efficiency and sub-nanosecond timing. The quenching of the Geiger discharge is either achieved by a resistor in series with each pixel or an active feedback.

Two types of SiPMs have been developed: Analogue and Digital. In Analogue SiPMs [47] the individual pixels are connected to a common readout and the SiPM delivers the summed analogue signal. In Digital SiPMs [48] each pixel has its own digital switch to a multi-channel readout system and the output is the digitized pulse height and precise time information for the pixels with Geiger discharges. Digital SiPMs also allow disabling pixels with high dark-count rates.

The pulse shape and the gain of SiPMs are explained with the help of Figs. 5.53 and 5.54: A schematic cross section of a single pixel is shown in Fig. 5.53, and an electrical model of a pixel with resistor quenching, in Fig. 5.54. The bias voltage is denoted V_{bias}, the single pixel capacitance C_{pix}, and the quenching resistance R_q. Frequently, in particular for SiPMs with larger pixel sizes, a capacitance C_q parallel to R_q is implemented. In the quiescent state the voltage over C_{pix} is V_{bias}. When an electron-hole pair in the amplification region starts a Geiger discharge, in the model the switch is closed and C_{pix} is discharged through the current source until the turn-off voltage V_{off} is reached, at which the Geiger discharge stops and the switch opens. The assumption of a constant current source is certainly oversimplified. However the sub-nanosecond discharge time is so short, that details of the time dependence of the discharge current hardly affect the results of the simulation. If a finite capacitance C_q is present, a fast pulse with charge $C_q \cdot (V_{bias} - V_{off})$ appears. After the switch opens, C_{pix} is charged up to V_{bias} with the time constant $\tau \approx R_q \cdot C_{pix}$ and the total signal charge is approximately $(C_{pix} + C_q) \cdot (V_{bias} - V_{off})$. Figures 5.55 and 5.56 show two examples of pulse shapes: (a) For a KETEK SiPM with $(15\ \mu m)^2$ pixels and negligible C_q, and (b) for a KETEK SiPM with similar doping profiles however with $(50\ \mu m)^2$ pixels and a finite C_q. The value of R_q has to be sufficiently high to quench the Geiger discharge. As C_{pix} increases with increasing pixel area, $\tau = R_q \cdot C_{pix}$ also increases, and a finite C_q has to be introduced to achieve a good timing performance and an increased pulse height if fast pulse shaping is used.

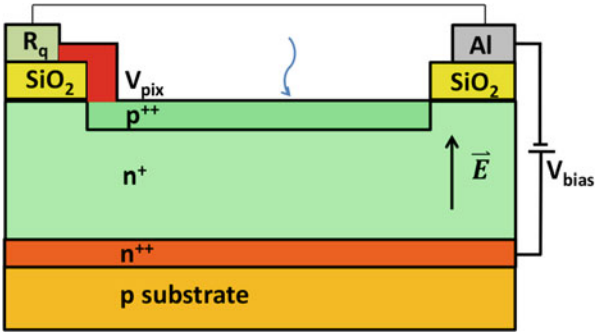


Fig. 5.53 Example of the schematic layout of a SiPM pixel. The Geiger breakdown occurs in the high-field n^+ region, which has a depth of order 1–2 μm. The p^{++}-electrode of every pixel is connected through the quenching resistance (R_q) to the biasing lines (Al) to which the biasing voltage V_{bias} is applied The photons enter through the transparent p^{++}-electrode

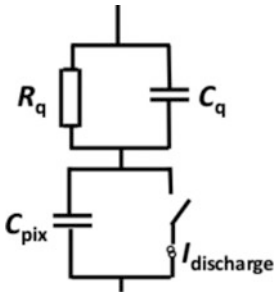


Fig. 5.54 Electrical model of a single SiPM pixel

Fig. 5.55 Pulse shape from a single photon for a KETEK SiPM with 4384 pixels of $(15\ \mu m)^2$: A single exponential with $\tau = R_q \cdot C_{pix} \approx 20$ ns

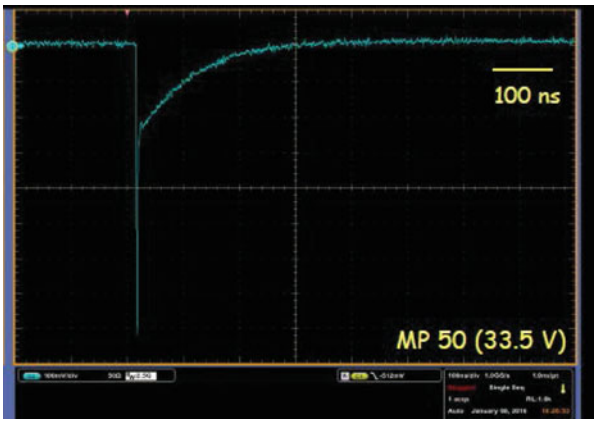


Fig. 5.56 Single photon pulse shape for a KETEK SiPM with 400 pixels of $(50\ \mu m)^2$: A prompt signal due to the finite value of C_q and a slow component with the time constant $\tau = R_q \cdot C_{pix} \approx 110$ ns is observed

In our discussion we distinguish between the breakdown voltage V_{bd}, the threshold voltage for a Geiger discharge, and the turn-off voltage V_{off}, the voltage at which the Geiger discharge stops. Differences $V_{bd} - V_{off}$ of up to about 1 V have been observed [49]. They should be taken into account when characterising or modelling SiPMs. We note that V_{bd} can be obtained from I–V measurements, as the voltage at which the current rises quickly due to the onset of Geiger discharges or the voltage at which the photon detection efficiency starts to differ from zero, and V_{off} can be determined from the dependence of SiPM Gain on V_{bias} by extrapolating the linear Gain(V_{bias}) dependence to Gain = 1.

One outstanding feature of SiPMs is the single-photon resolution, as demonstrated in the charge spectrum shown in Figs. 5.57 and 5.58 [50]. 0, 1, ... up to >30 simultaneous Geiger discharges can be distinguished allowing for straight-forward

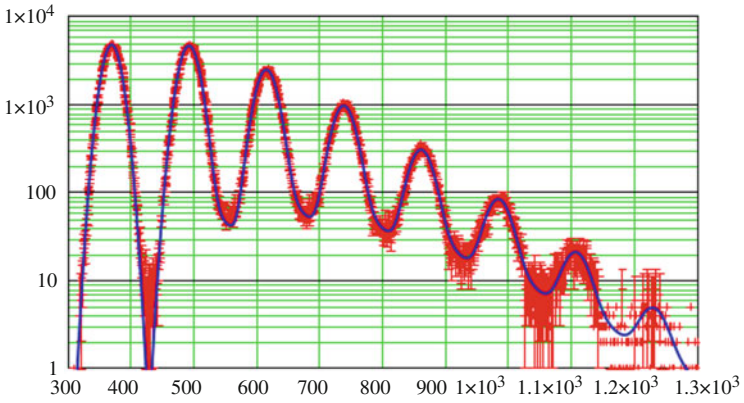


Fig. 5.57 Pulse height spectrum for a pulsed picosecond-laser measured with a KETEK SiPM with 4384 pixels of $(15\ \mu m)^2$. The solid curve is a model fit to the data. The average number of photons producing an initial Geiger discharge is 1.15

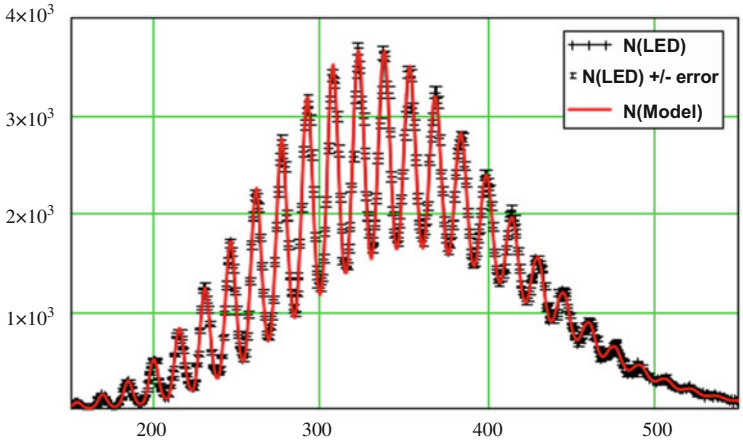


Fig. 5.58 Same as Fig. 5.57, however with an average number of photons producing an initial Geiger discharge of 18.6

calibration methods. The high photon-detection efficiency, where after careful optimisation values in excess of 60 % for wavelengths between 250 and 600 nm have been reached, the high gain of typically 10^6, and the intrinsic timing resolution of a few picoseconds, are other attractive performance parameters. In addition, SiPMs are not affected by magnetic fields, operate in a wide temperature range, are very robust, and work at moderate bias voltages ($\approx$ 25–75 V). Also, thanks to the microelectronics technology, SiPMs have highly reproducible performance parameters and are relatively inexpensive.

Limitations of SiPMs are their size, which is typically below 1 cm^2, and their limited dynamic range, essentially determined by the number of pixels. In addition, the measurement of the number of photons is affected by two sources of excess noise, which worsen the resolution beyond Poisson statistics: After-pulsing and Cross-talk. After-pulses are the result of charge carriers which are produced in the Geiger discharge and trapped in defect states. Depending on the energy in the silicon band gap and the properties of the defect states, they are released with different de-trapping time constants and cause additional signal fluctuations, which depend on the integration time of the readout electronics. In Figs. 5.57 and 5.58, which show pulse-height spectra recorded with a 100 ns gate at room temperature, after-pulses can be seen as entries in-between the peaks. Cross-talk is produced by the photons from the accelerated charges in the Geiger discharge, which generate electron-hole pairs in adjacent SiPM pixels. The photon path can be inside of the silicon but also via reflection in the protective layer of the SiPM or a light guide. This light path is so short that this cross-talk can be considered as prompt. Implementing trenches filled with absorbing material in-between the pixels reduces the prompt cross-talk significantly. The photons from the Geiger discharge can also generate electron-hole pairs in the non-depleted region of the SiPM, which can diffuse into

the amplification region and cause delayed cross-talk. The result of prompt cross-talk is that the number of entries in the peaks does not follow a Poisson distribution, even if the number of photons causing initial Geiger discharges does. As shown in [51] the result of cross-talk is that the number of entries in the peaks follows a Generalised-Poisson instead of a Poisson distribution. We note that the solid curve shown in Figs. 5.57 and 5.58 is the result of a model fit which includes both after-pulsing and prompt cross-talk simulated by a Generalised Poisson distribution. The model provides a fair description of the measurements and gives a precise determination of the SiPM parameters [50]. As both, after-pulses and cross-talk are related to the number of charge carriers in the Geiger discharge and thus to the Gain, the corresponding probabilities are expected to be approximately proportional to $V_{bias} - V_{off}$, which is also observed. Typical values at $V_{bias} - V_{off} = 5$ V for after-pulsing as well as prompt cross-talk are 5 % resulting in an excess noise factor, the ratio of the square of the relative resolution to the Poisson expectation, ENF $= [(\sigma_{meas}/mean_{meas})/(\sigma_{Poisson}/mean_{Poisson})]^2$ of ≈ 1.08. As the photon detection efficiency increases with voltage and finally saturates, whereas Gain and ENF continue to increase, there is a voltage at which the photon number measurement is optimal.

Dark counts are another limitation of SiPMs. Typical dark count rates (DCR) for SiPMs before irradiation are between 10 and 100 kHz/mm^2 at room temperature. Cooling reduces the DCR by about a factor 2 for an 8 °C reduction in temperature. Ionizing radiation, which mainly causes damage to the SiO_2, hardly affects the DCR. However non-ionizing radiation, like neutrons or high energy (> 5 MeV) particles, significantly affect the performance. At sufficiently high fluences (Φ) the DCR is so high that most pixels are in a state of Geiger discharge, the photon-detection efficiency decreases and finally the SiPM stops working as a photo-detector. Whereas V_{bd} and the electrical SiPM parameters hardly change up to $\Phi = 5 \times 10^{13}$ cm^{-2}, DCR increases by many orders of magnitude: For a KETEK SiPM with 15 μm pitch at –30 °C and $(V_{bias} - V_{off}) = 5$ V, DCR increases from ≈ 10 kHz/mm^2 before irradiation to ≈ 200 GHz/mm^2 after irradiation by reactor neutrons to $\Phi = 5 \times 10^{13}$ cm^{-2} [52, 53]. It is found that the increase in DCR is approximately proportional to Φ. It is also observed that after irradiation the increase of DCR with excess voltage is significantly steeper and the decrease with temperature slower after than before irradiation. As a result of the increased DCR, the signal baseline shows large fluctuations and single photon detection becomes impossible. Finally the occupancy of the pixels by dark counts is so high that the probability of a photon hitting a pixel which is already busy increases and the photon detection efficiency degrades. For the KETEK SiPM with 15 μm pitch at –30 °C the photon detection efficiency due to dark counts is reduced by a factor 2 for $\Phi = 5 \times 10^{13}$ cm^{-2} at $(V_{bias} - V_{off}) \approx 2.5$ V, and essentially zero for $\Phi = 5 \times 10^{14}$ cm^{-2} [53]. At these high fluences the dark currents exceed several tens of mA and thermal run-away has to be avoided.

After irradiation a significant reduction of DCR by annealing occurs. Annealing is a strong function of temperature: The typical reduction of DCR is a factor 2–3 after several days at room temperature, and a factor 10–50 at 175 °C. A systematic

study of different annealing scenarios, which allows to optimise the temperature cycling for operating SiPMs in high radiation fields, as available for silicon tracking detectors without gain [7, 10], is so far not available. In [54] it is demonstrated that SiPMs produced by Hamamatsu and SENSL, after irradiation to a fluence of 10^{14} cm^{-2} and annealed at 175 °C can achieve single photon detection at 77 K with a DCR below 1 kHz/cm^2.

The values of V_{bd} and V_{off} have a temperature dependence of order 20 mV/°C, which results in a temperature-dependent gain. However this is not a real problem and several feedback systems for gain stabilisations have been designed and are used.

Due to the vast application potential, which spans from research, over industrial applications to medicine, several firms develop and manufacture SiPMs. In close collaboration with research institutions, in particular working in particle physics, a rapid development and major improvements of SiPMs are presently under way.

5.11.4 Ultrafast Tracking Detectors: LGADs

At the HL-LHC (High-Luminosity Large Hadron Collider at CERN planned to start operation in 2026) in the large collider experiments ATLAS and CMS there will be on average ≈ 200 interactions with vertices distributed over ≈ 10 cm along the beam direction for every bunch crossing. For the complete kinematic reconstruction of the most interesting interactions in a bunch crossing, the information of the individual detector components has to be assigned to the correct interaction vertices. To illustrate the problem, Fig. 5.59 shows the reconstructed tracks extrapolated to the interaction region for a single bunch crossing with 50 interactions recorded in 2012. For a few vertices the interaction times, which are spread over ≈ ±200 ps, as obtained from a simulation, are given. For an efficient assignment of tracks to vertices, tracking detectors with high efficiency, 5 μm position resolution, 20 ps

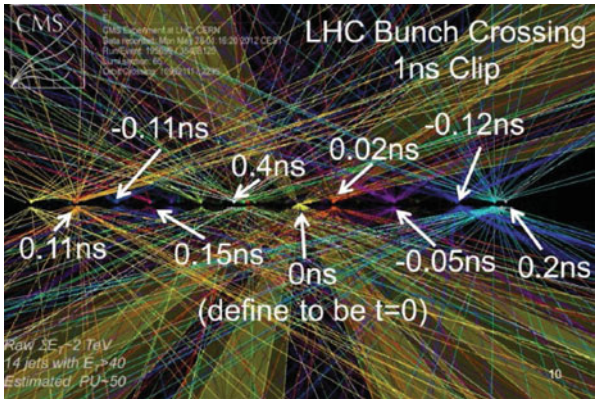


Fig. 5.59 Interaction times of a number of proton-proton vertices in a single bunch crossing with 50 interactions [55]. The data have been recorded by the CMS experiment in 2012. At the HL-LHC, the average number of interactions per bunch crossing is expected to be about 200

timing accuracy and 25 ns pulse shaping, are required. From simulations [55] it is concluded that pixel sensors with 50 μm active thickness and a doping profile similar to the one shown for APDs in Fig. 5.52 and operated at a gain of ≈ 20 can reach the required performance. These detectors are called Low-Gain-Avalanche Detectors, LGADs.

Different to optical photons which generate single electron-hole pairs, minimum-ionizing produce about 75 charge pairs per micro-meter and a high gain is not required. In addition to increasing fluctuations, high gain causes also practical difficulties and increases the shot noise from the dark current. Thin sensors have the additional advantage of smaller dark currents and a pulse rise time which increases with decreasing sensor thickness.

The effects which influence the timing accuracy can be grouped in five categories: (1) Position-dependent fluctuations of the charge carriers produced by the charged particle to be measured, (2) excess noise of the amplification mechanism, (3) position dependent drift field and coupling of the of the drifting charges to the readout electrodes, (4) electronics noise, and (5) digitisation error of the time-to-digital convertor.

A major issue for LGADs is the control of the gain after irradiation. The change of the effective doping by dopant removal and defect states, and the decrease of the mobilities and amplification coefficients of electrons and holes due to radiation damage appear to present major problems. These are addressed in an extensive R&D program which started in 2012 and has already given first encouraging results.

5.12 Summary and Outlook

Different concepts of solid silicon sensors and the electronics required for their readout have been described in this contribution. Although a detailed theoretical understanding of silicon devices had already been achieved in the 1960s, silicon detectors remained a niche application, used mainly in Nuclear Physics. This changed around 1980, when Josef Kemmer adapted the planar technology of micro-electronics to sensor fabrication and the ACCMOR Collaboration demonstrated the reliable long-term operation and excellent physics performance of silicon strip detectors. Based on these results, many groups started to develop and use silicon detectors, and today there is hardly a particle physics experiment, which does not rely heavily on them. The areas covered by silicon detectors in the particle physics experiments increased from tens of cm^2 to hundreds of m^2. Large areas of silicon detectors are even used on satellites for space experiments. In parallel to silicon detectors, the development of low-noise ASICs and connection technology started. They are required for reading out the more and more complex silicon sensors. In addition, a number of industrial producers, in closed collaboration with academia, developed and fabricated silicon sensors. Today silicon radiation detectors are a quite big market. Initially developed for Particle Physics, the use of silicon detectors spread into many different fields of science, medicine and industrial applications.

Since 1980 several new detector concepts were proposed, realised and used for a variety of measurement tasks. Outstanding examples are drift detectors, fully depleted CCDs, DEPFETs, MAPSs 3-D sensors, APDs and SiPMs. The different devices have their advantages and shortcomings, but offer high-performance solutions for most measurement tasks. In recent years radiation damage for the use of silicon sensors at high flux or high luminosity colliders has become more and more of a concern. Whereas radiation damage by X-rays can be controlled by a proper sensor design, the question up to which fluence of high-energy radiation silicon detectors can be used is a field of intense research. Unfortunately other sensor materials, like crystalline diamond or GaAs seem not to be a solution. Defect engineering, by doping crystals with different impurities has resulted in some improvements. However, a breakthrough for high fluences could not be demonstrated. Therefore the only approach appears to optimise the sensor layout for radiation tolerance. The recipe followed are high fields and low charge collection distances. How far intrinsic amplification can help remains an open question. For the design optimisation, complex TCAD (Technology Computer-Aided Design) simulations are performed. In spite of some first successes, a major progress is still required. As far as the electronics, which is exposed to the same fluences, is concerned, the sub-micron technology with nano-meter dielectric layers resulted in a big step in radiation tolerance.

For the future there is the strong hope that detectors can be fabricated which achieve the challenging performance parameters in the high radiation fields of the HL-LHC and future high-luminosity colliders. The field of solid state detectors will also profit very much from the ongoing industrial R&D efforts, in particular of 3-D integration technology and nano-electronics. Last but not least I very much hope that, like in the past, radically new ideas will come up and expand further the applications of solid state detectors.

References

1. E. Gatti, P. Rehak: *Semiconductor Drift Chamber - An Application of a Novel Charge Transport Scheme*, Nucl. Instrum. Meth. 225 (1984) 608-614; E. Gatti et al.: *Silicon Drift Chambers - First results and optimum processing of signals*, Nucl. Instrum. Meth. 226 (1984) 129-141; E. Gatti et al.: *Semiconductor Drift Chambers*, IEEE Trans. Nucl. Sci. 32 (1985) 1204-1208.
2. L. Strüder et al.: *The MPI/AIT X-ray imager (MAXI) - high speed pn-CCDs for X-ray detection*, Nucl. Instrum. Meth. A 288 (1990) 227-235; L. Strüder et al.: *First results with the pn-CCD detector system for the XMM satellite mission*, Nucl. Instrum. Meth. A 326 (1993) 129-135; L. Strüder et al.: *A 36 cm^2 large monolythic pn-charge coupled device X-ray detector for the European XMM satellite mission*, Rev. Sci. Instrum. 68 (1997) 4271-4274.
3. J. Kemmer, G. Lutz: *New semiconductor detector concepts*, Nucl. Instrum. Meth. A 253 (1987) 356-377.
4. J. Zhang et al.: *Study of radiation damage induced by 12 keV X-rays in MOS structures built on high-resistivity n-type silicon,* Journal of Synchrotron Radiation 19 (2012) 340-376.
5. T. Poehlsen, et al.: *Charge losses in segmented silicon sensors at the Si-SiO2 interface*, Nucl. Instrum. Meth. A 700 (2013) 22-39.

6. J. Schwandt et al.: *Design and First Tests of a Radiation-Hard Pixel Sensor for the European X-Ray Free-Electron Laser,* IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 61, NO. 4, AUGUST 2014 1894-1901.
7. M. Moll: *Displacement Damage in Silicon Detectors for High Energy Physics*, Manuscript accepted for Publication in IEEE Transactions on Nuclear Science, DOI:https://doi.org/10.1109/TNS.2018. 2819506.
8. V. Eremin, E. Verbitskaya, Z. Li: *The origin of double peak electric field distribution in heavily irradiated silicon detectors,* Nucl. Instrum. Meth. A 476 (2002) 556-564.
9. R. Klanner et al.: *Determination of the electric field in highly-irradiated silicon sensors using edge-TCT measurements,* Nucl. Instrum. Meth. A 951 (2020) 162987.
10. R. Wunstorf et al.: *Results on Radiation Hardness of Silicon Detectors up to Neutron Fluences of* $10^{15} n/cm^{-2}$, Nucl. Instrum. Meth. A 315 (1992) 149-155.
11. S.I Parker, C.J. Kenney and J. Segal: *3D - A proposed new architecture for solid state radiation detectors*, Nucl. Instrum. Meth. A 395 (1997) 328.
12. J. Kemmer et al.: *Experimental confirmation of a new semiconductor detector principle*, Nucl. Instrum. Meth. A 288 (1990) 92-98.
13. M. Caccia et al.: *A Si Strip Detector with Integrated Coupling Capacitors*, Nucl. Instrum. Meth. A 260 (1987) 124-131.
14. C. Cottini, E. Gatti, G. Gianelli, G. Rozzi: *Minimum noise preamplifiers for fast ionization chamber*, Nuovo Cimento (1956) 473-483.
15. E. Gatti, P.F. Manfredi: *Processing the signals from solid state detectors in elementary particle physics*, Rivista di Nuovo Cimento 9, Ser. 3 (1986) 1-145.
16. W. Buttler et al.: *Low-noise, low power monolithic multiplexing readout electronics for silicon strip detectors*, Nucl. Instrum. Meth. A 273 (1988) 778-783.
17. P. Jarron, et al.: *Deep submicron CMOS technologies for the LHC experiments*, Nucl. Phys. B Proc. Suppl. 78, no. 1-3 (1999) 625-634.
18. P. Rehak et al.: *Semiconductor drift chambers for position and energy measurements*, Nucl. Instrum. Meth A 235 (1985) 223-234.
19. A. Castoldi et al.: *A new drift detector with reduced lateral diffusion*, Nucl. Instrum. Meth A 377 (1996) 375-380.
20. W. Chen et al.: *Large area cylindrical silicon drift detector*, IEEE Trans. Nucl. Sci. 39 (1992) 619-628.
21. P. Rehak et al.: *Spiral silicon drift detectors*, IEEE Trans. Nucl. Sci. 36 (1989) 203-209.
22. R. Hartmann, et al.: *Ultrathin entrance windows for silicon drift detectors*, Nucl. Instrum. Meth. A 387 (1997) 250-254.
23. R. Hartmann et al.: *Design and test at room temperature of the first silicon drift detector with on-chip electronics*, IEDM Technical Digest (1994) 535-539.
24. V. Radeka et al.: *Implanted silicon JFET on Completely Depleted High Resistivity Devices*, IEEE El. Dev. Lett. 10, nb. 2 (1989) 91-95; E. Pinotti et al.: *The pn-CCD On-Chip Electronics*, Nucl. Instrum. Meth. A 326 (1993) 85-92.
25. R. Bailey et al.: *First Measurements of Efficiency and Precision of CCD Detectors for High Energy Physics*, Nucl. Instrum. Meth. 213 (1983) 201-215; C.J.S. Damerell et al.: *CCDs for Vertex Detection in High Energy Physics*, Nucl. Instrum. Meth. A 253 (1987) 478-481; C.J.S. Damerell et al.: *A CCD based vertex detector for SLD*, Nucl. Instrum. Meth. A 288 (1990) 236-239.
26. K. Abe et al.: *Design and performance of the SLD vertex detector: a 307 Mpixel tracking system*, Nucl. Instrum. Meth. A 400 (1997) 287-343.
27. D.H. Lumb et al.: *X-ray Multi-Mirror Mission - an overview*, SPIE 2808 (1997) 326-337.
28. G. Richter et al.: *ABRIXAS, A Broadband Imaging X-ray All-sky Survey*, (L. Bassani, G. di Cocco, eds.): Imaging in High Energy Astronomy, Experim. Astron. (1996) 159.
29. N. Meidinger et al.: *The PN-CCD detector for XMM and ABRIXAS*, SPIE 3765 (1999) 192-203.
30. L. Strüder et al.: *pnCCDs on XMM-Newton – 42 months in orbit*, Nucl. Instrum. Meth. A 512 (2003) 386-400.

31. L. Strüder et al.: *Large format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources*, Nucl. Instrum. Meth A614 (2010) 483-496.
32. B. Henrich et al.: *The adaptive gain integrating pixel detector AGIPD: A detector for the European XFEL*, Nucl. Instrum. Meth. A 633 (2011) S11-S14; A. *Allahgholi et al.*: *The adaptive gain integrating pixel detector*, JINST 11 (2016) C02066.
33. B. Henrich et al.: *PILATUS: A single photon counting pixel detector for X-ray applications*, Nucl. Instrum. Meth. A 607 (2009) 247-249.
34. R. Ballabriga, M. Campbell, X. Llopart, *ASIC Developments for radiation imaging applications: The medipix and timepix family*, Nucl. Instrum. Meth. A 878 (2018) 10-23.
35. W. Dulinski et al.: *Beam telescope for medium energy particles based on thin, submicron precision MAPS*, in: Nuclear Science Symposium Conference Record, 2007, NSS '07, IEEE, Vol. 2, 995-1002.
36. L. Ratti et al., *CMOS MAPS with fully integrated, hybrid-pixel-like analog front-end electronic*, eConf C0604032 (2006) S.0008.
37. I. Peric, *A novel monolithic pixelated particle detector implemented in high-voltage CMOS technology*, Nucl. Instrum. Meth. A 582 (2007) 876-885; I Peric et al.: *High-voltage pixel detectors in commercial CMOS technologies for ATLAS, CLIC and Mu3e experiments*, Nucl. Instrum. Meth. A 731 (2013) 131-136.
38. XEUS Astrophysics working group: *X-ray Evolving - Universe Spectroscopy - The XEUS scientific case*, ESA SP-1238 (1999), 30 pages.
39. S. Wölfel et al.: *Sub electron noise measurements on repetitive non-destructive readout devices*, Nucl. Instrum. Meth. A 566 (2006) 536-539.
40. G. Lutz, R.H. Richter, L. Strüder: *Halbleiterstruktur, insbesondere in einem Halbleiterdetektor, und zugehöriges Betriebsverfahren*, EU Patent 1 873 834; G. Lutz, et al.: *DEPFET detector-amplifier structure: Properties, achievements and new developments, concepts and applications*, in: Nuclear Science Symposium Conference Record, 2007, NSS '07, IEEE, Vol. 2, 988-994.
41. G. Lutz, L. Strüder: *DEPFET Transistor mit großem Dynamikbereich und Halbleiterdetektor*, DE Patent 10 2007 048 890; G. Lutz et al.: *DEPFET Sensor with intrinsic signal compression developed for use at the XFEL free electron laser radiation source*, Nucl. Instrum. Meth. A 624 (2010) 528-532.
42. J. Treis et al,: *DEPFET based instrumentation for the MIXS focal plane on BepiColombo*, in: Instrumentation and Methods for Astrobiology and Planetary Missions XII (R.B. Hoover, G.V. Levin, A. Yu Rozanov, K. Retherford, eds.), Proc. SPIE 7441 (2009) 774116.
43. M. Porro et al.: *Performance of ASTEROID: A 64 channel ASIC for source follower readout of DEPFET matrices for X-ray astronomy*, IEEE Nuclear Science Symposium Conference Record 2008, pp. 1830-1835.
44. R.H. Richter et al.: *Design and technology of DEPFET pixel sensors for linear collider applications*, Nucl. Instrum. Meth. A511 (2003) 250-256.
45. L. Andricek et al.: *The MOS-type DEPFET pixel sensor for the ILC environment*, Nucl. Instrum. Meth. A 565 (2006) 165-171.
46. L. Andricek, G. Lutz, M. Reiche, R.H. Richter: *Processing of ultra-thin silicon sensors for future e^+e^- linear collider experiments*, IEEE Trans. Nucl. Sci. 51 (2004) 1117-1120.
47. D. Renker: *Geiger-Mode Avalanche Photodiodes, history properties and problems*, Nucl. Instrum. Meth. A 567 (2006) 48-56; D. Renker and E. Lorenz: *Advances in solid state photon detectors*, JINST 4 (2009) P04004.

48. T. Frach et al.: *The Digital Silicon Photomultiplier – Principle of Operation and Intrinsic Detector Performance*, in: Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), N34-4 (2012) 1959-1965; C. Degenhardt et al.: *The digital Silicon Photomultiplier – A novel sensor for the detection of scintillation light*, in Proceeding to IEEE NSS-MIC conference, Orlando U.S.A. October 25–31 2009, Proc. IEEE 2009 (2009) 2383-2386; S. Mandai, E. Charbon: *Multi-channel digital SiPMs: concept, analysis and implementation*, in: Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), N34-4 (2012) 1840-1844.
49. V. Chmill et al.: *Study of the breakdown voltage of SiPMs*, Nucl. Instr. Meth. A 845 (2017) 56-59.
50. V. Chmill, et al.: *On the characterisation of SiPMs from pulse-height spectra*, Nucl. Instr. Meth. A 854 (2017) 70-81.
51. S. Vinogradov: *Analytical models of probability distribution and excess noise factor of solid state photomultiplier signals with crosstalk*, Nucl. Instrum. Meth. A 695 (2012) 247-251.
52. Yu. Musienko et al.: *Radiation damage studies of silicon photomultipliers for the CMS HCAL phase I upgrade,* Nucl. Instr. Meth. A787 (2015) 319–322; Yu. Musienko et al.: *Effects of very high radiation on SiPMs*, Nucl. Instr. Meth. A824 (2016) 111-114.
53. M. Centis Vignali et al.: *Neutron irradiation effect on SiPMs up to* $\Phi = 5\text{x}10^{14}$ cm^{-2}, Nucl. Instr. Meth. A912 (2018) 137.
54. M. Calvi et al.: *Single photon detection with SiPMs irradiated up to* 10^{14} cm^{-2} *1-MeV-equivalent neutron fluence*, arXiv:1805.07154 (2018).
55. N. Cartiglia et al.: *Tracking in 4 dimensions*, Nucl. Instr. Meth. A845 (2017) 47-51.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 6
Calorimetry

Particle Detectors and Detector Systems

C. W. Fabjan and D. Fournier

6.1 Introduction, Definitions

In particle physics, calorimetry refers to the absorption of a particle and the transformation of its energy into a measurable signal related to the energy of the particle. In contrast to tracking a calorimetric measurement implies that the particle is completely absorbed and is thus no longer available for subsequent measurements.

If the energy of the initial particle is much above the threshold of inelastic reactions between this particle and the detector medium, the energy loss process leads to a cascade of lower energy particles, in number commensurate with the incident energy. The charged particles in the shower ultimately lose their energy through the elementary processes mainly by ionization and atomic level excitation. The neutral components in the cascade (γ, n,..) contribute through processes described later in this section.

The sum of the elementary losses builds up the calorimetric signal, which can be of ionization or of scintillation nature (or Cherenkov) or sometimes involve several types of response.

While the definition of calorimetry applies to both the low energy case (no showering) and the high energy case (showering), this section deals mostly with the showering case. Examples of calorimetry without showering are discussed in Sect. 6.2.3.

C. W. Fabjan
Austrian Academy of Sciences and University of Technology, Vienna, Austria
e-mail: Chris.Fabjan@cern.ch

D. Fournier (✉)
IJCLab, Université Paris-Saclay, CNRS/IN2P3, Orsay, France
e-mail: daniel.fournier@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_6

Only electromagnetic and strong interactions contribute to calorimetric signals, the weak (and gravitational) interaction being much too small to contribute. Particles with only weak (or gravitational interaction) will escape direct calorimetric detection. An exception are the neutrino detectors discussed in Sect. 6.4: statistically, when a very large number of neutrinos cross a detector, a tiny fraction of them will interact (weakly) with matter and will lead to particle production which can be measured by different methods, including calorimetry.

The measurement of the energy of a particle is the primary goal of calorimetry. In addition, several other important quantities can be extracted, such as impact position and timing, particle direction and identification. These issues are considered in Sects. 6.4–6.6, before addressing specific examples in Sect. 6.7.

In Sect. 6.2 the fundamentals of calorimetry are presented, followed by a discussion of signal formation obtained from the energy deposition (Sect. 6.3).

In recent years, calorimetry in the ATLAS and CMS detectors at the LHC played an essential role in the discovery of the Higgs boson, announced in July 2012.

6.2 Calorimetry: Fundamental Phenomena

Given the large differences between electromagnetic interactions and strong interactions, the following subsections start with electrons and photons, which have only electromagnetic interactions (see however the end of this section), before addressing the case of particles with strong interactions, also called hadrons. The case of muons is considered in a separate subsection.

6.2.1 Interactions of Electrons and Photons with Matter

Several elementary interaction processes of the electrons with the medium contribute to the energy loss $-\mathrm{d}E$ of an electron of energy E after a path dx in a medium: Møller scattering, ionization and scattering off the nuclei of the medium: bremsstrahlung (Fig. 6.1). Electron-electron scattering is considered as ionization (Møller) if the energy lost is smaller (larger) than $m_e c^2/2$. It is customary to include

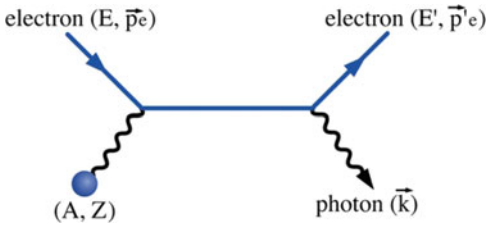


Fig. 6.1 Photon radiation from electron interaction with a nucleus (A, Z)

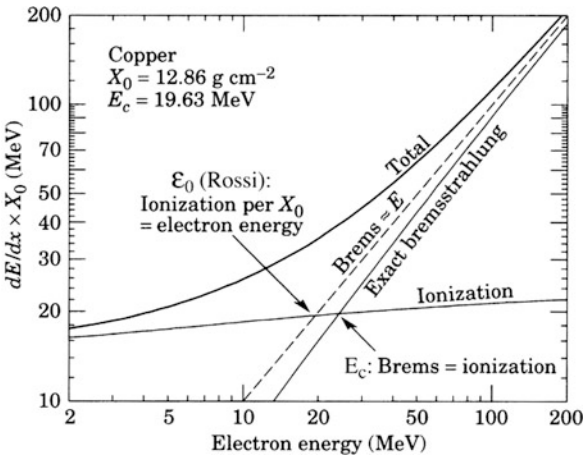


Fig. 6.2 Average energy loss of electrons in copper by ionization and bremsstrahlung. Two definitions of the critical energy (E and ε (Rossi)) are shown by arrows

in energy loss by ionization atomic excitations, some of which lead to light emission (scintillation). For positrons the Møller scattering is replaced by Bhabha scattering.

The calculated average energy loss is shown in Fig. 6.2 for copper and the average fractional energy loss ($-1/E\ \mathrm{d}E/\mathrm{d}x$) is plotted in Fig. 6.3 for lead [1].

Figure 6.2 illustrates that the average energy lost by electrons (and positrons see Fig. 6.3) by ionization is almost independent of their incident energy (above ~1 MeV), with however a small logarithmic increase. For electrons [1, 2].

$$\frac{-\mathrm{dE}}{\mathrm{dx}} = \mathrm{k}\frac{\mathrm{Z}}{\mathrm{A}}\frac{1}{\beta^2}\left\{\ln\frac{\gamma \mathrm{m_e c^2}\beta\sqrt{\gamma-1}}{\mathrm{I}\sqrt{2}} + \frac{1}{2}\left(1-\beta^2\right) - \frac{2\gamma-1}{2\gamma^2}\ln 2 + \frac{1}{16}\left(\frac{\gamma-1}{\gamma}\right)^2\right\}\left(\mathrm{MeV}/\left(\mathrm{g/cm^2}\right)\right) \tag{6.1}$$

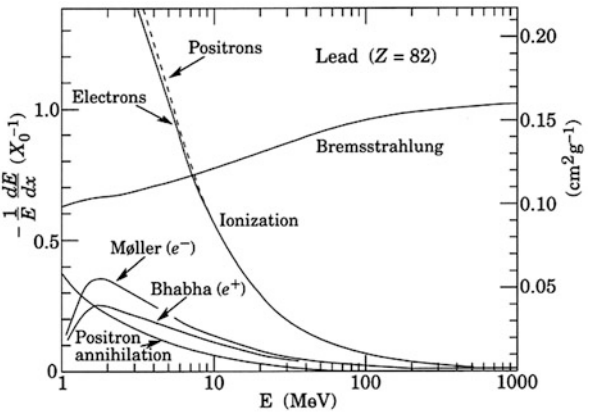


Fig. 6.3 Relative energy loss of electrons and positron in lead with the contributions of ionization, bremsstrahlung, Møller (e^-) and Bhabha (e^+) scattering and positron annihilation

and for positrons

$$-\frac{\mathrm{dE}}{\mathrm{dx}} = \mathrm{k}\frac{\mathrm{Z}}{\mathrm{A}}\frac{1}{\beta^2}\left[\ln\frac{\gamma \mathrm{m_e c^2}\beta\sqrt{\gamma-1}}{\mathrm{I}\sqrt{2}} - \frac{\beta^2}{24}\left(23+\frac{14}{\gamma+1}+\frac{10}{(\gamma+1)^2}+\frac{4}{(\gamma+1)^3}\right)\right]\left(\mathrm{MeV}/\left(\mathrm{g/cm^2}\right)\right) \tag{6.2}$$

In these formula, A (Z) are the number of nucleons (protons) in the nuclei of the medium, I is the mean excitation energy of the medium—often approximated by 16 $Z^{\mathbf{0.9}}$ eV—the constant $k = 4\pi N_\mathrm{A} {r_\mathrm{e}}^2 m_\mathrm{e} c^2 = 0.3071$ MeV/(g/cm^2), N_A the Avogadro number and $\mathrm{r_e} = \frac{1}{4\pi\varepsilon_0}\cdot\frac{\mathrm{e}^2}{\mathrm{m_e c^2}} = 2.818\ 10^{-15}$ m the classical radius of the electron.

For positrons the annihilation with an electron of the medium has to be considered. The cross section of this process ($\sigma_{\mathrm{an}} = Z\pi\,{\mathrm{r_e}}^2/\gamma$ for $\gamma >> 1$) decreases rapidly with increasing energy of the positron. At very low energy, the annihilation rate is:

$$R = NZ\ \pi\ {\mathrm{r_e}}^2\ c\ \left[\mathrm{s}^{-1}\right], \tag{6.3}$$

with $N = \rho\ N_\mathrm{A}/A$, the number of atoms per unit volume.

This rate corresponds to a lifetime in lead of about 10^{-10}s [3]. Positron annihilation plays a key role in some technical applications (Positron Emission Tomography, Chap. 7).

Figure 6.2 shows that the average energy loss by bremsstrahlung (photon emission in the electromagnetic field of a nucleus) increases almost linearly as a function of incident energy (meaning that the fractional energy loss is almost constant, as shown in Fig. 6.3).

This is described by introducing the *radiation length* X_0 defined by:

$$-\mathrm{d}E/E = \mathrm{d}x/X_0 \tag{6.4}$$

It follows from the definition that X_0 is the mean distance after which an electron has lost, by radiation, all but a fraction 1/e of its initial energy. X_0 also has a simple meaning in terms of photon conversion (see below).

While X_0 should show a small increase at low energy corresponding to a small drop in the fractional energy loss visible in Fig. 6.3, it soon reaches a high-energy limit which has been calculated by Bethe and Heitler [3, 4] and more recently by Tsai [5] and tabulated by Dahl [1] for different materials. In the seminal book by Rossi [6] the formula for X_0, based on the Bethe–Heitler formalism reads:

$$1/X_0 = 4\,\alpha\ (N_\mathrm{A}/A)\ \left\{Z\,(Z+1)\,{r_\mathrm{e}}^2 \ln\left(183 Z^{-1/3}\right)\right\}\ \left[\mathrm{cm^2\ g^{-1}}\right] \tag{6.5}$$

The Z^2 term reflects the fact that the bremsstrahlung results from a coupling of the initial electron to the electromagnetic field of the nucleus, somewhat screened by the electrons (*log* term), and augmented by a direct contribution from the electrons (Z^2 replaced by $Z\,(Z+1)$).

The radiation length of a compound, or mixture, can be calculated using:

$$1/X_0 = \Sigma\ w_\mathrm{j}/X_\mathrm{j} \tag{6.6}$$

where the w_j are the fractions by weight of the nuclear species j of the mixture or of the compound.

The spectrum of photons with energy k radiated by an electron of energy E traversing a thin slab of material (expressed as a function of $y = k/E$) has the characteristic "bremsstrahlung" spectrum:

$$\mathrm{d}\sigma/\mathrm{d}k = A/\left(X_0 N_\mathrm{A} k\right) \cdot \left(4/3 – 4/3\mathrm{y} + \mathrm{y}^2\right). \tag{6.7}$$

At very high energies a number of effects, considered at the end of this subsection, modify the spectrum.

Another important quantity, the *critical energy* can be introduced examining Fig. 6.2. The critical energy E_c for *electrons* (or *positrons*) in a given medium is defined as the energy at which energy loss by radiation in a thin slab equals the energy loss by ionization. A slightly different definition ε_0, introduced by Rossi, results from considering the relative energy loss as fully independent of energy (see Fig. 6.2). The critical energy ε_0 is well described in dense materials (see Fig. 6.4) by:

$$\varepsilon_0 = 610\ \mathrm{MeV}/\left(Z + 1.24\right). \tag{6.8}$$

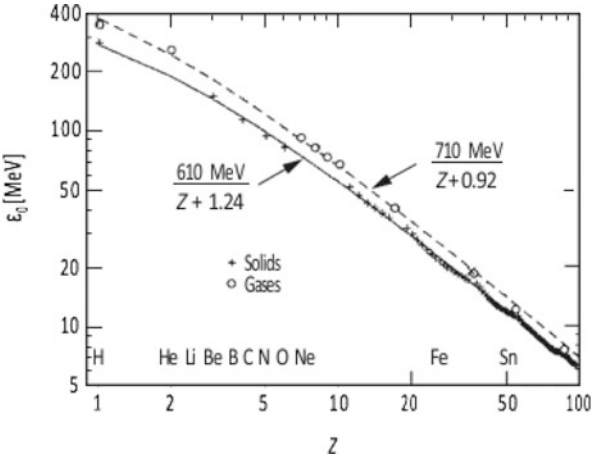


Fig. 6.4 Critical energy for the chemical elements, using Rossi's definition [6]. The fits shown are for solids and liquids (solid line) and gases (dashed line)

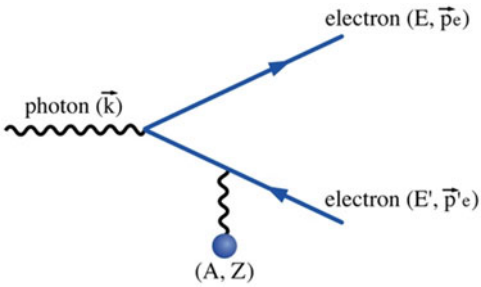


Fig. 6.5 Electron-positron pair creation in the field of a nucleus (A, Z)

As will be seen below, X_0 and E_c (or ε_0) are among the important parameters characterizing the formation of electromagnetic showers.

Several processes contribute to the interaction of photons with matter, the relative importance of which depends primarily on their energy.

Pair Production

This process is dominant as soon as photon energies are above a few times 2 $m_e c^2$. The graph responsible of the process (Fig. 6.5) shares the vertices of the bremsstrahlung graph.

The dominant part (Z^2) is due to the nucleus, while the electrons contribute proportionally to Z. The process of pair production has been studied in detail [7]. The pair production cross section can be written, in the complete screening limit at high energy as:

$$\mathrm{d}\sigma/\mathrm{d}x = A/(X_0 N_\mathrm{A}) \cdot (1\text{–}4/3x\ \ (1-x))\,, \tag{6.9}$$

where $x = E/k$ is the fraction of the photon energy k taken by the electron of the pair. Integrating the cross section over E gives the pair production cross section:

$$\sigma = 7/9\ A/(X_0 N_\mathrm{A})\,. \tag{6.10}$$

After 9/7 of an X_0, the probability that a high-energy photon survives without having materialized into an electron-positron pair is 1/e. In the pair production process the energy of the recoil nucleus is small, typically of the order of $m_e c^2$, implying that at high photon energy ($k >> m_e c^2$) the electron and the positron are both collinear with the incident photon. When the reaction takes place with an electron, the momentum transfer can be much higher leading to "triplets" with one positron and two electrons in the final state.

As for bremsstrahlung the cross section is affected at very high energy by processes considered later.

Compton Effect

The QED cross-section for the photon-electron scattering (Klein-Nishina [8]) can be written in the limit of $k >> m_e c^2$, using $x = k/m_e c^2$,

$$\sigma = \pi {r_e}^2\,(\log\ 2x + 1/2)/x\ \left[\mathrm{cm}^2\right]. \tag{6.11a}$$

The related probability for Compton scattering after the traversal of a material slab of thickness dt and mass per unit volume ρ is:

$$\phi = \sigma\rho \; N_{\mathrm{A}} \; Z/A \; \mathrm{d}t. \tag{6.11b}$$

For high Z (e.g. lead) the maximum of the Compton cross section and the pair production cross-section are of the same order of magnitude, while for lighter materials the maximum of the Compton cross section is higher. This is illustrated in Fig. 6.6 (from [1]) where carbon and lead are compared.

The differential Compton cross-section, with θ denoting the scattering angle between the initial and final photon, and η the angle between the vector perpen-

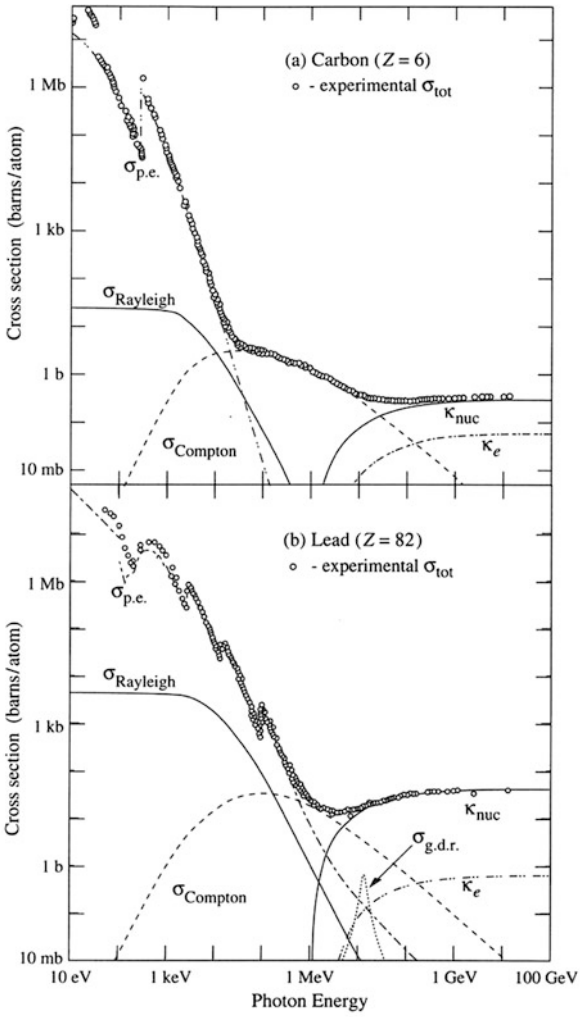


Fig. 6.6 Photon total cross section as a function of the photon energy in carbon and lead, with the contributions of different processes. $\sigma_{\mathrm{p.e.}}$ corresponds to the atomic photoelectric effect and κ_{nuc} (κ_{e}) corresponds to pair production in the nuclear (electron) field

dicular to the scattering plane and the polarization vector of the initial photon (in case it is linearly polarized) reads (ε being the ratio between the scattered and the incident photon energy $\varepsilon = 1/(1 + k/m_e c^2(1 - \cos\theta))$:

$$\mathrm{d}\sigma/\mathrm{d}\Omega = 0.5\, r_e^2 \left(\varepsilon + 1/\varepsilon - 2\sin^2\theta\cos^2\eta\right). \tag{6.12}$$

At low energy (k not larger than a few MeV), the η-dependence can be exploited for polarization measurements (Compton polarimetry). In the same energy range the probability of backward scattering is also sizeable.

Photoelectric Effect

For sufficiently low photon energies the atomic electrons can no longer be considered as free. The cross section for photon absorption, followed by electron emission (photoelectric effect) presents discontinuities whenever the photon energy crosses the electron binding energy of a deeper shell.

Explicit calculations [4] show that above the K-shell the cross section decreases like $E^{-3.5}$.

In the section devoted to shower formation, the relevance of the photoelectric effect will be considered. The coherent scattering (or Rayleigh scattering) is comparatively smaller than the photoelectric effect and its role negligible for shower formation.

High Energy Effects (LPM)

In the collinear approximation of bremsstrahlung, the longitudinal momentum difference $q_{||}$ between the initial electron (energy E) and the sum of the final electron and photon (energy k) is equal to

$$q_{||} = {m_e}^2 c^3 k/2E\,(E - k)\,. \tag{6.13}$$

This quantity can be extremely small, being for example 0.002 eV/c for a 25 GeV electron radiating a 10 MeV photon. Such a small longitudinal momentum transfer implies a large formation length, L_f ($L_f\ q_{||} \geq h/2\pi$), about 100 μm in the above example. Secondary interactions (like multiple scattering) taking place over this distance will perturb the final state and will in general diminish the bremsstrahlung cross section and the pair production cross section in case of photon interactions. Coherent interaction of the produced photons with the medium (dielectric effect) also affects, and reduces, the bremsstrahlung cross-section.

Such effects, already anticipated by Landau and Pomeranchuk [9] were considered in detail by several authors, and were measured by the experiment E146 at SLAC. A recent overview is given in [10]. The high k/E part of the bremsstrahlung spectrum is comparatively less affected (because of much larger $q_{||}$ values) while the low k/E part is significantly influenced for E above ~100 GeV, see Fig. 6.7. Only at much higher energies (>10 TeV) is the pair production cross-section affected.

In crystalline media the strong intercrystalline electrical fields may result in coherent suppression or enhancement of bremsstrahlung. Net effects depend on the

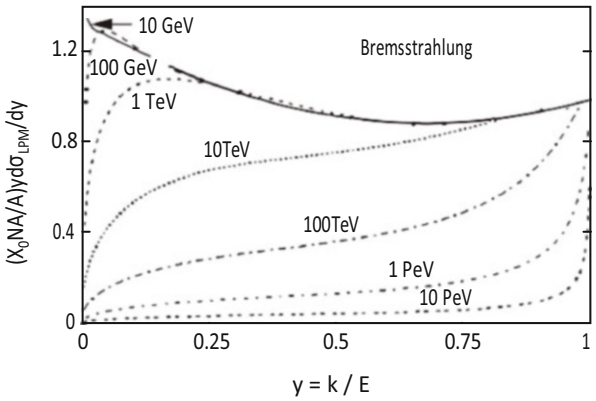


Fig. 6.7 Normalized Bremsstrahlung cross-section $k\,\mathrm{d}\sigma/\mathrm{d}k$ in lead as a function of the fraction of momentum taken by the radiated photon

propagation direction of the particle with respect to the principal axes of the crystal [11].

Hadronic Interactions of Photons

Photons with energies above a few GeV can behave similarly to Vector Mesons (ρ, ω and ϕ) with the same quantum numbers and in this way develop strong interactions with hadronic matter. They can be parameterized with the Vector Meson Dominance model. Using the Current-Field Identity [12], the amplitude for interactions of virtual photons γ^* of transverse momentum q is:

$$\mathcal{A}\left(\gamma^* \mathrm{A} \to \mathrm{B}\right) = \left(e/2\gamma_\rho\right)\ \mathrm{m}^2{}_\rho/\left(\mathrm{m}^2{}_\rho - q^2\right)\mathcal{A}\left(\rho \mathrm{A} \to \mathrm{B}\right) \\ +\text{equiv.terms for } \omega \text{ and } \phi \text{ mesons.} \tag{6.14}$$

Various photo- and electro-production cross sections were calculated and confronted with experiment. As an example the ratio of hadron production to electron-positron pair production in the interaction of a 20 GeV photon is about 1/200 for hydrogen and 1/2500 for lead [13]. While this ratio is small, the effect on shower characteristics and on particle identification can in certain cases be significant [for example—see Ref. 14—when studying CP violating $\pi\pi$ final states in $\mathrm{K_L}$ decays, for which $\pi e \nu$ decays are a background source].

6.2.2 Electromagnetic Showers

When a high energy electron, positron or photon impinges on a thick absorber, it initiates an electromagnetic cascade as pair production, bremsstrahlung and Compton effects generate electrons/positrons and photons of lower energy. Electron/positron energies eventually fall below the critical energy, and then dissipate their energy

by ionization and excitation rather than by particle production. Photons propagate somewhat deeper into the material, being ultimately absorbed primarily via the photoelectric process.

Given the large number of particles (electrons, positrons, photons) present in a high energy electromagnetic cascade (more than one thousand for a 10 GeV electron or photon in lead), global variables have been sought to describe the average shower behaviour. Scale variables, such as X_0 as unit length, can be used to parameterize the radiation effects. However, since energy losses by $\mathrm{d}E/\mathrm{d}x$ and by radiation depend in a different way on material characteristics, one should not expect perfect 'scaling'.

Analytical Description

In an analytical description [6] a first simplification consists in 'factorizing' the longitudinal development and the lateral spread of showers, with the assumption that the lateral excursion of electrons and photons around the direction of the initial particle does not affect the longitudinal behaviour and in particular the 'total track length' (see below).

As for any statistical process the first goal is to obtain analytical expressions for average quantities. Particularly relevant (for a shower of initial energy E_0) are: $c(E_0,E,t)$ the average number of electrons plus positrons with energy between E and $E + \mathrm{d}E$ at depth t (expressed in radiation length), and the integral distribution $C(E_0,E,t) = \int_0^E c(E_0,E',t)\mathrm{d}E'$; $n(E_0,E,t)$ and $N(E_0,E,t)$ are the corresponding functions for photons.

Using the probability distribution functions of the physical effects driving the shower evolution (Bremsstrahlung, Compton, $\mathrm{d}E/\mathrm{d}x$, pair production) one can write and solve [15, 16] 'evolution equations' correlating $C(E_0,E,t)$ and $N(E_0,E,t)$. In the so called 'approximation B' of Rossi, the energy loss of electrons by $\mathrm{d}E/\mathrm{d}x$ is taken as constant, and the pair production and bremsstrahlung cross-sections are approximated by their asymptotic expression.

As an illustration, Fig. 6.8 shows the number of electrons and positrons as a function of depth, in a shower initiated by an electron of energy E_0, and by a photon of energy E_0 in units of the "Rossi critical energy ε_0" (see Sect. 6.2.1). These distributions are integrated over E from 0 to the maximum possible. The area under the curves is to a good approximation equal to E_0/ε_0, in accordance with the physical meaning of ε_0. The two sets of curves also show that a photon initiated shower is shifted on average by about 1 X_0 to larger depths compared to an electron (or positron) initiated one.

The total track length $TTL = \int_0^\infty C(E_0,0,t)\mathrm{d}t$ the energy transferred to the calorimeter medium by $\mathrm{d}E/\mathrm{d}x$, the source of the calorimeter signal.

Results from Monte Carlo Simulations

While analytical descriptions are useful guidelines, many applications require the use of Monte-Carlo (MC) simulations reproducing step by step, in a statistical manner, the physical effects governing the shower formation. For several decades, the standard simulation code for electromagnetic cascades has been EGS4 [17]. A recent alternative is encoded in the Geant4 framework [18].

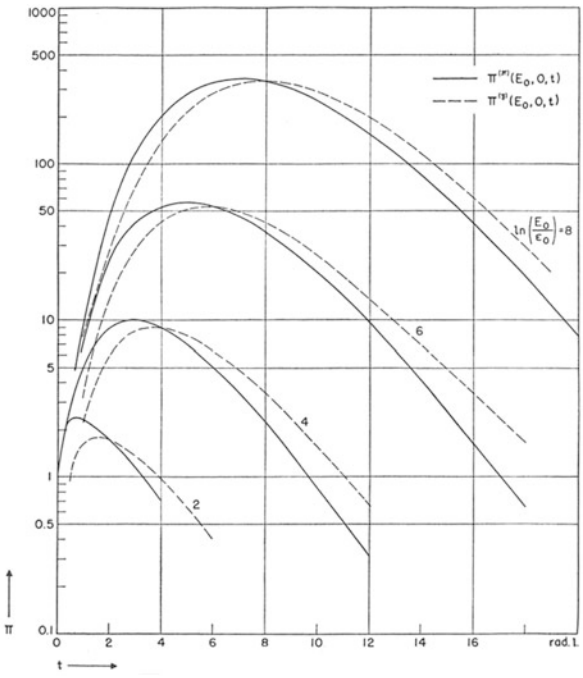


Fig. 6.8 Number of charged secondaries as a function of shower depth, for an electron initiated shower (full lines) and a photon initiated one (dashed lines), calculated analytically by Snyder [15, 16]. The numbers attached to each set of curves indicate $\ln(E_0/\varepsilon_0)$

As an illustration of the additional information obtained by this MC approach, Fig. 6.9 shows results of a 30 GeV electron shower simulation in iron ($E_c = 22$ MeV). The energy deposition per slab ($dt = 0.5X_0$) is shown as a histogram, with the fitted analytical function (see below) superimposed. This distribution is close, but not identical, to the distribution of electrons above a certain threshold (here taken as 1.5 MeV) crossing successive planes (right-hand scale): the energy deposition is slightly below the number of electrons at the beginning of the shower, and somewhat higher at the end. Multiple scattering (see below), affecting more the low energy shower tail, is one effect contributing to this discrepancy. The distribution of photons above the same threshold of 1.5 MeV is shifted to larger X_0 with respect to the electron distribution, reflecting the higher penetration power of photons already mentioned.

As a further illustration of the power of MC simulations, Fig. 6.10 displays longitudinal profiles of 10 GeV electron showers obtained by Geant4 simulation in lead, copper and aluminium. Since the dE/dx per X_0 is relatively more important in low Z material compared to high Z materials, one expects showers to penetrate more deeply in high Z materials, a fact born out by the simulations. Illustrating the energy dependence of shower parameters Fig. 6.11 displays shower energy deposition as a function of depth (shower profiles) for a range of incident electron energies (1 GeV to 1 TeV) in lead. The position of the shower maximum shows

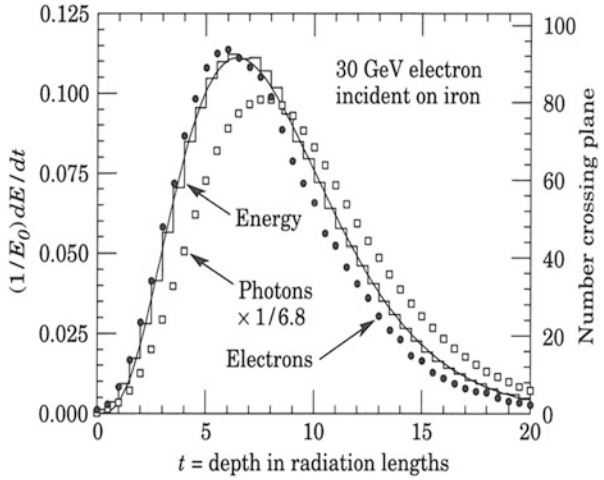


Fig. 6.9 EGS4 simulation of a 30 GeV electron-induced cascade in iron. The histogram is the fractional energy deposition per radiation length, and the curve is a gamma function fit to the distribution. The full (open) points represent the number of electrons (photons) with energy greater than 1.5 MeV crossing planes at $X_0/2$ intervals

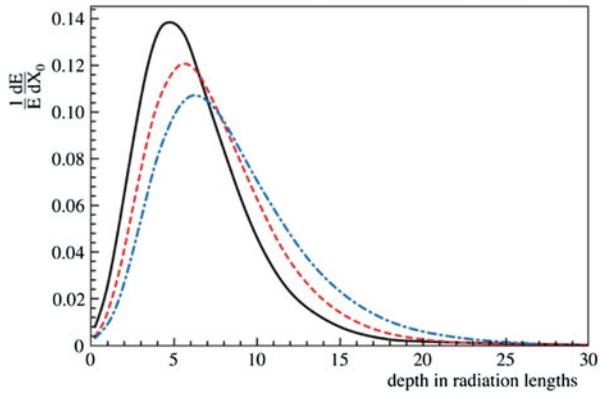


Fig. 6.10 Fractional energy deposition per longitudinal slice of 1 X_0 for 10 GeV electrons in aluminium (full line), copper (dashed) and lead (dash-dotted) (Geant4)

the expected logarithmic dependence on incident energy. In the parameterisation of shower profiles by Longo and Sestili [19].

$$F(E,t) = E_0\, b(bt)^{a-1}\, \mathrm{e}^{-bt}/\Gamma(a) \tag{6.15}$$

one finds accordingly $t_{\max} = (a-1)/b$, well fitted by $t_{\max} = \ln(y) + C_{\mathrm{i}}$, ($C_{\mathrm{i}} = 0.5$ for photons, -0.5 for electrons, and $y = E/E_{\mathrm{c}}$).

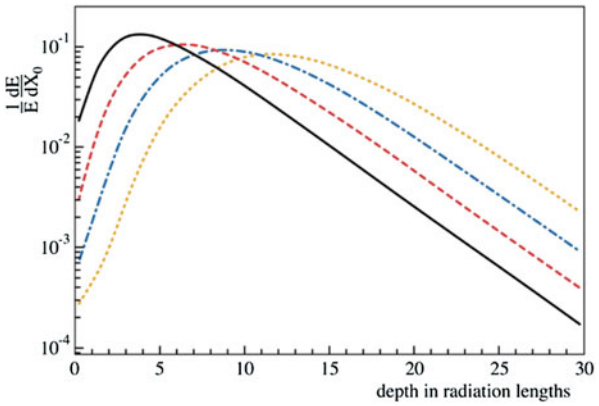


Fig. 6.11 Fractional energy deposition in lead, per longitudinal slice of 1 X_0, for electron induced showers of 1 GeV (full line), 10 GeV (dashed), 100 GeV (dash-dotted) and 1 TeV (dotted) (Geant4)

Finally, Fig. 6.12 illustrates the imbalance between electrons and positrons: in an electromagnetic shower, and rather material independent, about 75% of the energy deposited by charged particles is due to electrons, and 25% to positrons. This imbalance is due to the Compton and photoelectric effects which generate only electrons. It is more important towards the end of the shower.

Lateral Shower Development

Bremsstrahlung and pair creation on nuclei take place without appreciable momentum transfer to the (heavy) nuclei. Bremsstrahlung on electrons of the medium and Compton scattering involve however some momentum transfer. For example, in the Compton interaction of a 2 MeV (0.5 MeV) photon, 6% (16%) of the scattered photons are emitted with an angle larger than 90° with respect to the initial photon direction z. Another important effect contributing to the transverse spread in a cascade is multiple scattering of electrons and positrons.

After a displacement of length l along z, in a medium of radiation length X_0, the projected rms angular deviation along the transverse directions x and y, of an electron of momentum p is:

$$\theta_{x,y} = \frac{E_s}{\sqrt{2}} \frac{1}{p\beta c} \sqrt{l/X_0} \tag{6.16}$$

and the lateral displacement is

$$\delta_{x,y} = \frac{\theta_{x,y} l}{\sqrt{3}} \tag{6.17}$$

with $E_s = m_e c^2 \sqrt{(4\pi/\alpha)} = 21.2$ MeV. The lateral displacement contributes directly to the transverse shower broadening. If, after a step of length l, the electron emits

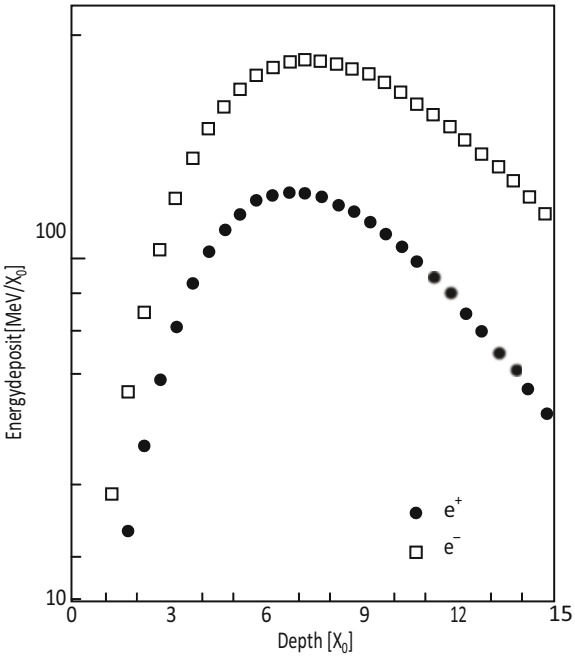


Fig. 6.12 Energy deposited in longitudinal slices of 1 X_0 by electrons (open symbols) and positrons (closed symbols) in a 10 GeV electron shower developing in lead (EGS4)

a bremsstrahlung photon, the emission will take place along the direction of the electron after l, thus at some angle (rms $\theta_{x,y}$ in both directions) with respect to the initial electron. Since the photon travels on average a considerable distance before materializing (9/7 X_0 if the photon is above a few MeV, significantly more at lower energy, see Fig. 6.6), the angular deviation of the electron gives a second, large contribution to the shower broadening.

In order to quantify the transverse shower spread, it is customary to use as parameter the Molière radius defined as:

$$\rho_{\mathrm{M}} = E_{\mathrm{s}}\, X_0 / E_{\mathrm{c}}, \tag{6.18}$$

where ρ_{M} equals $\sqrt{6}$ times the transverse displacement of an electron of energy E_{c}, after a path (without radiation nor energy loss) of 1 X_0. The most relevant physical meaning of ρ_{M} comes from Monte-Carlo simulations which show that about 87% (96%) of the energy deposited by electrons/positrons in a shower is contained in a cylinder of radius 1 (2) ρ_{M}.

Going back to the expressions of X_0 and E_{c}, it can be seen that their ratio is proportional to A/Z, and thus ρ_{M} is rather independent from the nuclear species, and is essentially governed by the material density. Calculations of ρ_{M}, for some pure materials and mixtures are reported in Table 6.1.

Table 6.1 Properties of calorimeter materials

Material	Z	Density [g cm^{-3}]	E_c [MeV]	X_0 [mm]	ρ_M [mm]	λ_{int} [mm]	$(dE/dx)_{mip}$ [MeV cm^{-1}]
C	6	2.27	83	188	48	381	3.95
Al	13	2.70	43	89	44	390	4.36
Fe	26	7.87	22	17.6	16.9	168	11.4
Cu	29	8.96	20	14.3	15.2	151	12.6
Sn	50	7.31	12	12.1	21.6	223	9.24
W	74	19.3	8.0	3.5	9.3	96	22.1
Pb	82	11.3	7.4	5.6	16.0	170	12.7
^{238}U	92	18.95	6.8	3.2	10.0	105	20.5
Concrete	–	2.5	55	107	41	400	4.28
Glass	–	2.23	51	127	53	438	3.78
Marble	–	2.93	56	96	36	362	4.77
Si	14	2.33	41	93.6	48	455	3.88
Ge	32	5.32	17	23	29	264	7.29
Ar (liquid)	18	1.40	37	140	80	837	2.13
Kr (liquid)	36	2.41	18	47	55	607	3.23
Polystyrene	–	1.032	94	424	96	795	2.00
Plexiglas	–	1.18	86	344	85	708	2.28
Quartz	–	2.32	51	117	49	428	3.94
Lead-glass	–	4.06	15	25.1	35	330	5.45
Air 20°, 1 atm	–	0.0012	87	304 m	74 m	747 m	0.0022
Water	–	1.00	83	361	92	849	1.99

mip minimum-ionizing particle

Comparing as an illustration lead and copper, one observes that the transverse dimensions of showers expressed in mm are essentially the same (because the transverse profiles are almost identical expressed in ρ_M (Fig. 6.14) and the ρ_M's are similar), while the shower in copper is (in mm) a factor 2.5 longer (because X_0 (copper) = 14.3 mm against 5.6 mm for lead).

On the other hand, despite being much shorter (in mm), the shower in lead contains about 2.5 more electrons (of lower energy in average) than the shower in copper, in the inverse proportion to their respective critical energies (7.4 MeV for lead against 20 MeV for Copper).

The lateral spread of showers is on average narrow at the beginning, where the shower content is still dominated by particles of energy much larger than E_c. In the low-energy tail the shower broadens. Monte Carlo simulations allow studying profiles at various depths. This is illustrated in Fig. 6.13 which shows the 90% containment radius as a function of the shower depth and in Fig. 6.14 which shows the radial profile of showers in three different materials. The broader width in the first 2 or 3 X_0 can be associated with backscattering (albedo) from the shower, which competes with the narrow core of the shower in its very early part. There is almost

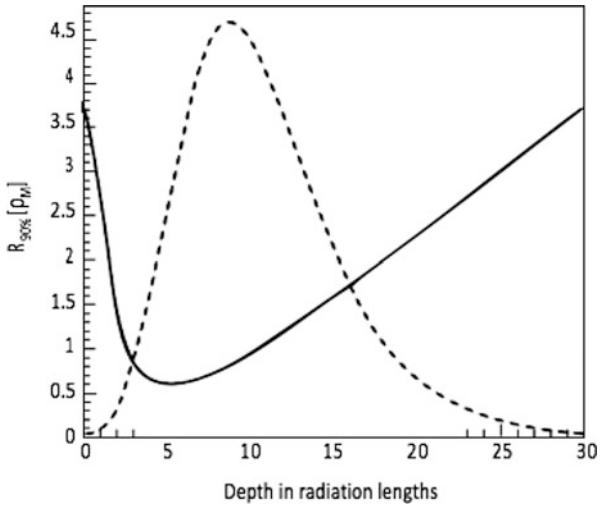


Fig. 6.13 90% containment radius $R_{90\%}$ (full line), in Molière radius ρ_M as a function of shower depth, for 100 GeV electron showers developing in lead. For comparison the longitudinal energy deposition is also shown (dashed line, arbitrary scale) (Geant4)

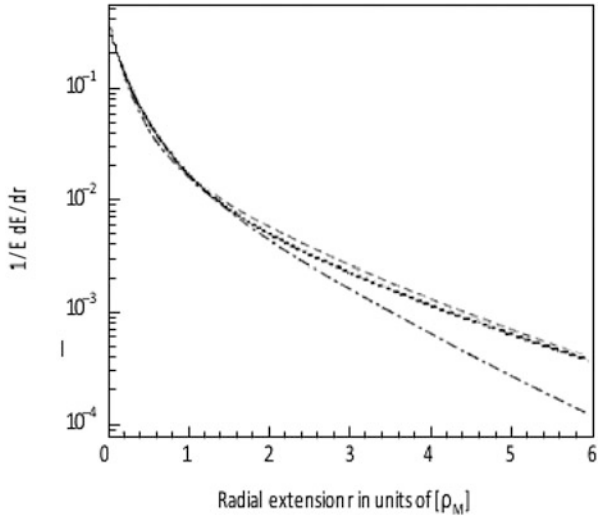


Fig. 6.14 Fractional energy deposition in cylindrical layers of thickness 0.1 ρ_M, coaxial with the incident particle direction, for 100 GeV electron-induced showers in aluminium (dotted line), copper (dashed line) and lead (dash-dotted) (Geant4)

no dependence of shower transverse profiles (integrated over depth) as a function of initial electron energy.

6.2.3 Homogeneous Calorimeters

For reasons explained later, large calorimeter systems are often 'sampling' calorimeters. These calorimeters are built as a stack of passive layers, in general of high Z material for electromagnetic calorimeters, alternating with layers of a sensitive medium responding to ('sampling' the) electrons/positrons of the shower, produced mostly in the passive layers.

A homogeneous calorimeter is built only from the sensitive medium. Provided all other conditions are satisfied (full containment of the shower, efficient collection and processing of the signal) homogeneous calorimeters give the best energy resolution, because sampling calorimeters are limited by 'sampling fluctuations' (see Sect. 6.2.4). It is instructive to study first the limitations in the "ideal" conditions of homogeneous calorimeters.

We first discuss low-energy applications, where the absorption does not involve showering. As an illustration, Fig. 6.15 shows the extremely narrow lines observed [20] when exposing a Germanium (Li-doped) crystal to a γ source of ^{108m}Ag and ^{110m}Ag. The resolution, at the level of one part in a thousand, is far better than obtained with NaI, a frequently used scintillating crystal (see below). Several

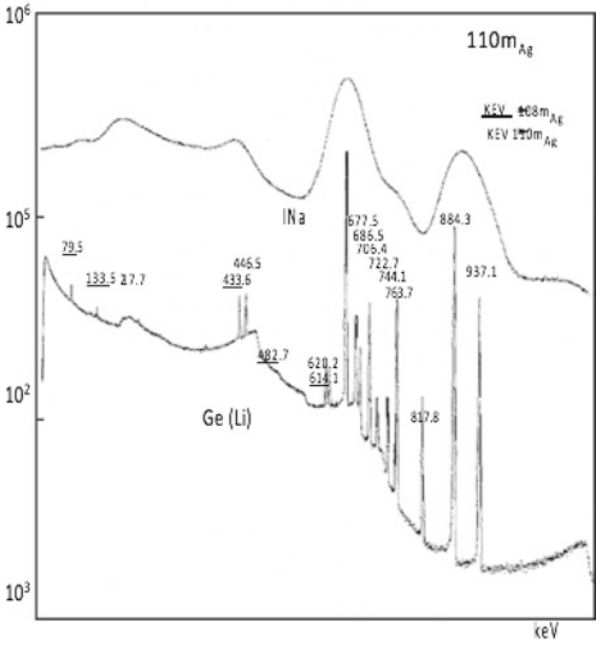


Fig. 6.15 Pulse height spectra recorded using a sodium iodide scintillator and a Ge (Li) detector. The source is a gamma radiation from the decay of ^{108m}Ag and ^{110m}Ag. Energies of peaks are labelled in keV

quantitative studies of the energy resolution of high purity Ge crystals, operated at low temperature (77 K) for γ spectroscopy have been made. A rather comprehensive discussion is given in [21]. After subtraction of the electronics noise, the width of the higher energy lines (above 0.5 MeV) is narrower than calculated assuming statistical independence of the created electron-hole pairs (~2.9 eV are needed to create such a pair). The reason for this was first understood by Fano [22]. Fundamentally it is due to the fact that the pairs created are not statistically independent, but are correlated by the constraint that the total energy loss must precisely be equal to the energy of the incident photon (in the limit of a device in which all energy losses lead to a detected signal, in a proportional way, the line width vanishes).

Calling σ the rms of the energy ε used to create an electron-hole pair, the actual resolution should be $\sigma/(\varepsilon\sqrt{Np})$, smaller than $1/\sqrt{Np}$ by a factor $\sqrt{F}$, where $F = (\sigma/\varepsilon)^2$ is the Fano factor. Monte-Carlo simulations [23] reproduce the phenomenon and give $F \sim 0.1$ for semiconductor devices, in reasonable agreement with measurements [21].

When two energy loss mechanisms compete, e.g. ionization and scintillation, the total energy constrain remains, but with a binomial sharing between the two mechanisms. It is thus expected that summing up the two contributions, assumed to be read out independently, will lead to an improved energy resolution (it should be remembered however that a certain fraction of the energy lost in the medium goes to heat.

This was first demonstrated with a liquid argon gridded cell exposed to La ions with an energy of 1.2 GeV/nucleon traversing the cell [24]. In this set-up both scintillation photons and electrons from electron-ion pairs were detected (see Sect. 6.3.3 for the collection mechanism). More recently, detailed studies of scintillation and ionization yields were made in liquid xenon using 662-keV γ-rays from a ^{137}Cs source [25]. With decreasing voltage applied to the sensitive liquid Xe volume, the scintillation signal increases while the ionization one decreases, as expected from recombination of electrons-ions giving rise to additional photons. The spectra obtained with scintillation alone, ionization alone, and their sum are shown in Fig. 6.16, together with the correlation between the two signals.

The ratio between scintillation and ionization depends also on the nature and energy of the particle making the deposit. Low energy nuclear recoils are highly ionizing, giving rise to more recombination and thus an increased light over charge ratio.

As discussed in Sect. 6.3.1, noble liquid detectors (using either argon or xenon) have been developed in the last decade which allowed pushing the limits of dark matter searches. They rely heavily on the existence of two correlated signals (ionization and scintillation) for a given energy deposit, exploiting in particular the ratio between the two to distinguish nuclear recoils from photon or muon background (see Sect. 6.7.2).

When the energy loss per unit length becomes very high (i.e. for low values of β and/or high values of the electric charge Ze for ions) saturation effects are observed in liquid ionization detectors, and also in scintillators. Empirically, the

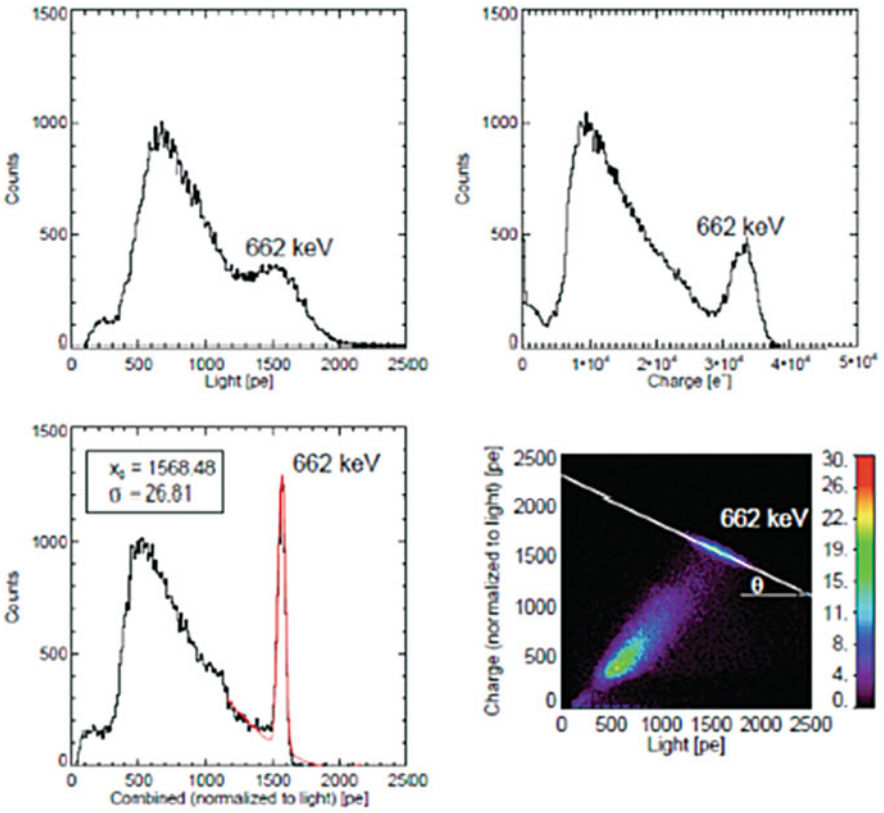


Fig. 6.16 Correlation between scintillation and ionization signals [25]. Scintillation alone (top-left), ionization alone (top-right), sum of both (bottom-left), 2-D correlation between scintillation and ionization (bottom-right)

effective scintillation (ionization) signal dL/dx (dI/dx) can be parameterized with "Birks law" [26]:

$$dL/dx = L_0.dE/dx/(1. + k_B\ dE/dx), \tag{6.19}$$

in which L_0 is the luminescence at low specific ionization density. The effect in plastic scintillators, for which k_B ~0.01 g.cm^{-2} MeV^{-1}, results in suppression ("quenching") of the light emission by the high density of ionized and excited molecules. Deviations from Birk's law have been observed for high Z ions [27].

In liquid ionization detectors the effect is associated with electron-ion recombination. It depends upon the electric field, in magnitude and direction with respect to the ionizing track. A typical value in liquid argon is k_B ~0.04 g.cm^{-2} MeV^{-1} for an electric field in a direction perpendicular to the track of 1 kV/cm, with k_B being inversely proportional to E, for E < 1 kV/cm [28].

Saturation effects are not relevant for electron or photon induced showers (at least below few TeVs) because the track density remains comparatively low (however, depending on the technique used for sampling calorimeters, internal amplification—like in calorimeters with gaseous readout—may saturate for high track density). Saturation effects do affect hadronic showers because of slow, highly ionizing fragments from nuclear break-up and slow proton recoils.

The—in general excellent—energy resolution of homogeneous calorimeters used for electromagnetic showers is affected by several instrumental effects. One of the most fundamental ones, the existence of a threshold energy E_{th} above which an electron of the shower does produce a signal will be illustrated in Sect. 6.3.2 when dealing with Cherenkov based electromagnetic calorimeters. Other effects include:

- longitudinal and transverse shower containment
- efficiency of light collection
- photoelectron statistics
- electron carrier attachment (impurities)
- space charge effects, . . .

These effects will be considered when dealing with examples where they are particularly relevant. The closer a detector approaches the intrinsic resolution—like for Ge crystals—the more important are the above limitations. In practice, large calorimeter systems for high energy showers based on homogeneous semiconductor crystals are unaffordable. Scintillating crystals and pure noble liquids are the best compromise between performance and cost, but do suffer from other limitations, as illustrated in examples given below.

6.2.4 Sampling Calorimeters and Sampling Fluctuations

In the simplest geometry, a sampling calorimeter consists of plates of dense, passive material alternating with layers of sensitive material.

For electromagnetic showers, passive materials with low critical energy (thus high Z) are used, thus maximizing the number of electrons and positrons in a shower to be sampled by the active layers. In practice, lead is most frequently used. Uranium has also been used to optimize the response towards hadrons (Sect. 6.2.7), and tungsten has been used in cases where compactness is a premium.

The thickness t of the passive layers (in units of X_0) determines the sampling frequency, i.e. the number of times a high energy electron or photon shower is 'sampled'. Intuitively, the thinner the passive layer (i.e. the higher the sampling frequency), the better the resolution should be. The thickness u of the active layer is usually characterized by the *sampling fraction* f_S which is the ratio of dE/dx of a

minimum ionizing particle in the active layer to the sum of $\mathrm{d}E/\mathrm{d}x$ in the active and passive layers:

$$f_S = u\ \mathrm{d}E/\mathrm{d}x_{\mathrm{active}}/\left(u\ \mathrm{d}E/\mathrm{d}x_{\mathrm{active}} + t\ \mathrm{d}E/\mathrm{d}x_{\mathrm{passive}}\right)\left[u, t \text{ in g cm}^{-2}, \mathrm{d}E/\mathrm{d}x \text{ in MeV/g cm}^{-2}\right]. \tag{6.20}$$

This 'sampling' of the energy results in a loss of information and hence in additional 'sampling fluctuations'. An approximation [29, 30] for these fluctuations in electromagnetic calorimeters can be derived using the total track length (*TTL*) of a shower initiated by an electron or photon of energy E. The signal is approximated by the number N_x of e^+ or e^- traversing the active signal planes, spaced by a distance $(t + u)$. This number N_x of crossings is

$$N_x = TTL/(t+u) = E/(\varepsilon_0\ (t+u)) = E/\Delta E,$$

ΔE being the energy loss in a unit cell of thickness $(t + u)$. Assuming statistical independence of the crossings, the fluctuations in N_x represent the 'sampling fluctuations' $\sigma(E)_{\mathrm{samp}}$,

$$\begin{aligned}\sigma(E)_{\mathrm{samp}}/E &= \sigma\left(N_x\right)/N_x = 1/\sqrt{N_x} = \sqrt{\{\Delta E\ (\mathrm{GeV})\ /E\ (\mathrm{GeV})\}} \\ &= 0.032\sqrt{\{\Delta \mathrm{E}\ (\mathrm{MeV})\ /\mathrm{E}\ (\mathrm{GeV})\}} = a/\sqrt{E}.\end{aligned} \tag{6.21}$$

The detector dependent constant a is the 'sampling term' of the energy resolution (see also below). For illustration, for a lead/scintillator calorimeter with 1.4 mm lead plates, interleaved with 2 mm scintillator planes, $\Delta E = 2.2$ MeV, one estimates $a \sim 5\%$ for 1 GeV electromagnetic showers. This represents a lower limit (the experimental value is closer to 7 to 8%), as threshold effects in signal emission and angular spread of electrons around the shower axis worsen the resolution [29]. In addition, a large fraction of the shower particles are produced as e^+e^- pairs, reducing the number of statistically independence crossings N_x.

The sampling fraction f_S has practical consequences, considering the actual signal produced by the calorimeter. If f_S is too small, the signal is small and may be affected by electronics noise and possibly other technical limitations due to the chosen readout technique (see below).

The dominant part of the calorimeter signal is actually not produced by minimum ionizing particles, but rather by the low-energy electrons and positrons crossing the signal planes. Defining the fractional response f_R of a given layer "i" as the ratio of energies lost in the active layer and of the sum of active plus passive layers one has

$$f_R^i = E^i_{\mathrm{active}}/(E^i_{\mathrm{active}} + E^i_{\mathrm{passive}}) \tag{6.22}$$

with the constraint that $\Sigma^i\ (E^{\,i}{}_{\mathrm{active}} + E^{\,i}{}_{\mathrm{passive}}) = E$.

Experimentally one finds that f_R (taking all layers together) is significantly smaller than f_S [31]. The ratio f_R/f_S, usually called '*e/mip*' for obvious reasons, can

be as low as 0.6 when the Z of the passive material (lead) is much larger than the Z of the active one (plastic scintillator, liquid argon). This effect, well reproduced by Monte-Carlo simulations, is to some extent due to the "transition effect" between the passive and active material, but also due to the fact that a significant fraction of electrons produced in the high Z passive material by pair production or Compton scattering do not have enough energy to exit this layer and are thus not sampled. This same effect induces a depth dependence of *e/mip*, which decreases by few percent towards the end of the shower.

Taking into account an energy independent contribution from electronics noise b, and a minimum asymptotic value of the relative energy resolution c (constant term, due for example to inhomogeneities in materials, imperfection of calibrations, ...) the energy resolution of a sampling calorimeter is in general written as[1]

$$\Delta E/E = a/\sqrt{E} \oplus b/E \oplus c \tag{6.23}$$

Experimentally it has been observed that the same relation holds also for homogeneous calorimeters, in general with smaller 'sampling terms' a, although their origin is not coming from sampling fluctuations, but from other limitations (see Sect. 6.2.3).

6.2.5 *Physics of the Hadronic Cascade*

By analogy with electromagnetic showers, the energy degradation of high-energy hadrons proceeds through an increasing number of (mostly) strong interactions with the calorimeter material. However, the complex hadronic and nuclear processes produce a multitude of effects that determine the performance of practical instruments, making hadronic calorimeters more complicated instruments to optimize and resulting in a significantly worse intrinsic resolution compared to the electromagnetic one. Experimental studies by many groups helped to unravel these effects and permitted the design of high-performance hadron calorimeters.

The hadronic interaction produces two classes of secondary processes. First, energetic secondary hadrons are produced with momenta typically a fair fraction of the primary hadron momentum, i.e. at the GeV scale. Second, in hadronic collisions with the material nuclei, a significant part of the primary energy is consumed by nuclear processes such as excitation, nucleon evaporation, spallation, etc., generating particles with energies characteristic of the nuclear MeV scale.

The complexity of the physics is illustrated in Fig. 6.17, which shows the energy spectra of the major shower components (weighted by their track length in the shower) averaged over many cascades, induced by 100 GeV protons in lead. These spectra are dominated by electrons, positrons, photons, and neutrons at low energy.

[1] In a formula like (6.23), $a \oplus b$ means $\sqrt{(a^2 + b^2)}$.

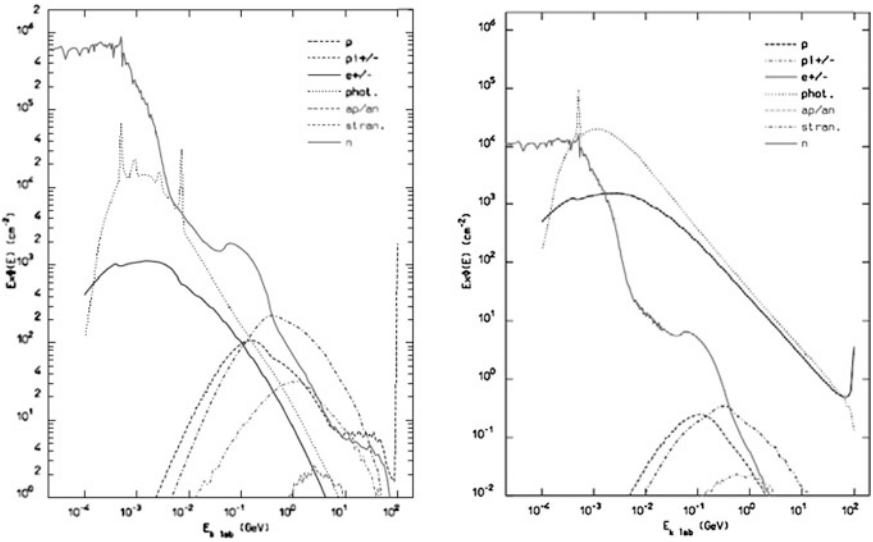


Fig. 6.17 Particle spectra produced in the hadronic cascade initiated by 100 GeV protons absorbed in lead (left). The energetic component is dominated by pions, whereas the soft spectra are composed of photons and neutrons. The ordinate is in 'lethargic' units and represents the particle track length, differential in log E. The integral of each curve gives the relative fluence of the particle [32]. On the right, same figure for 100 GeV electrons in lead, showing the much simpler structure, dominated by electrons and photons (hadrons are down by more than a factor 100)

The structure in the photon spectrum at approximately 8 MeV reflects a (n,γ) reaction and is a fingerprint of nuclear physics; the line at 511 keV results from e^+e^- annihilation photons. These low-energy spectra encapsulate all the information relevant to the hadronic energy measurement. Deciphering this message becomes the story of hadronic calorimetry.

The energetic component contains protons, neutrons, charged pions and photons from neutral pion decays. Due to the charge independence of hadronic interactions, on average approximately one third of the pions produced will be π^0s, $f_{\pi 0} \approx 1/3$. These neutral pions will decay to two photons, $\pi^0 \rightarrow \gamma\gamma$, before reinteracting hadronically and will induce an electromagnetic cascade, proceeding along its own laws of electromagnetic interactions (see Sect. 6.2.2). This physics process acts like a 'one-way diode', transferring energy from the hadronic part to the electromagnetic component, which will not contribute further to hadronic processes.

As the number of energetic hadronic interactions increases with increasing incident energy, so will the fraction of the electromagnetic cascade. This simple picture of the hadronic showering process leads to a power law dependence of the two components [33, 34]; naively, the electromagnetic component is $F_{\mathrm{em}} = 1 - (1 - f_{\pi 0})^n$, n denoting the number of shower generations induced by a particle with energy E. For the hadronic fraction F_{h} one finds in a more realistic evaluation $F_{\mathrm{h}} = (E/E_0)^k$. The parameter k expresses the energy dependence and is related to the average

multiplicity m of a collision, with $k = \ln(1 - f_{\pi 0})/\ln m$. The parameter E_0 denotes the average energy necessary for the production of a pion, approximately $E_0 \approx 2$ GeV; with the multiplicity $m \approx 6$–7 of hadrons produced in a hadronic collision k is ≈ -0.2. Values of F_h are of order 0.5 (0.3) for 100 (1000) GeV showers. As the energy of the incident hadron increases, it is doomed to dissipate its energy in a flash of photons. Were one to extrapolate this power law to the highest particles energies detected calorimetrically, $E \leq 10^{20}$ eV more than 98% of the hadronic energy would be converted to electromagnetic energy!

The low-energy nuclear part of the hadronic cascade has very different properties, but carries the dominant part of the energy in the hadronic sector. In the energetic hadron collisions with the nuclei of the calorimeter material, their nucleons will be struck initiating an 'intra-nuclear' cascade. In the subsequent steps, the intermediate nucleus will de-excite, in general through a spallation reaction, evaporating a considerable number of nucleons, accompanied by few MeV γ-emission. The binding energy of these nucleons released in these collisions is taken from the energy of the incident hadron. The number of these low-energy neutrons is large: ~ 20 neutron/GeV in lead. The fraction of the total associated binding energy depends on the incident energy and may be as high as ~20–40%. These neutrons will ultimately be captured by the target nuclei, resulting in delayed nuclear photon emission (at the ~ μs timescale). The energy lost to binding energy is therefore, in general, not detected ('invisible') in practical calorimeters.

In Fig. 6.18 the energy dependence of the electromagnetic, fast hadron and nuclear components is shown. The response of a calorimeter is determined by the sum of the responses to these different components which react with the passive and

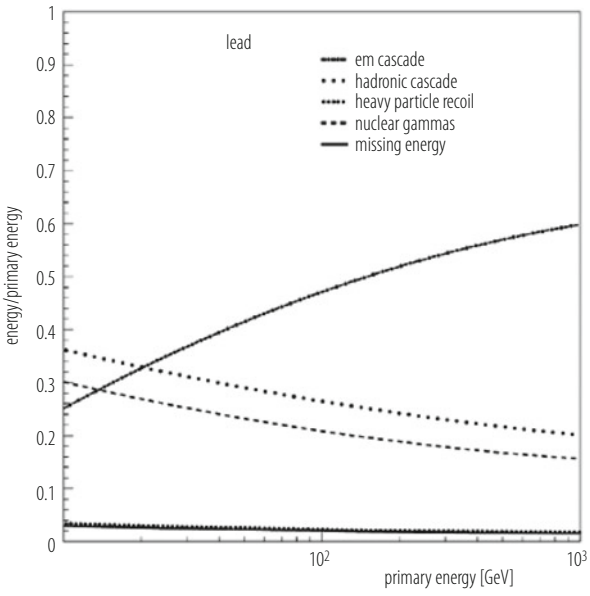


Fig. 6.18 Characteristic components of proton-initiated cascades in lead. With increasing energy the em component increases [32]

active parts of the calorimeter in their specific ways (see Sect. 6.2.7). Contributions from neutrons and photons from nuclear reactions, which have consequences for the performance of these instruments, are also shown in Fig. 6.18. The total energy carried by photons from nuclear reactions is substantial: only a fraction, however, will be recorded in practical instruments, as most of these photons are emitted with a considerable time delay (~1 μs). The event-by-event fluctuations in the invisible energy dominate the fluctuations in the detector signal, and hence the energy resolution. The road to high-performance hadronic calorimetry has been opened by understanding how to compensate for these invisible energy fluctuations [35].

6.2.6 *Hadronic Shower Profile*

The total cross section for hadrons is only weakly energy dependent in the range of few to several hundred GeV, relevant for calorimetry. For protons, the total pp. cross section σ_{tot} is approximately 39 mb. For pion-proton collisions $\sigma_{\text{tot}}(\pi\text{p}) = 2/3\ \sigma_{\text{tot}}(\text{pp})$ is naively expected, i.e. 26 mb, compared to the measured value of $\sigma_{\text{tot}}(\pi^{+}\text{p}) \approx 23$ mb. For hadronic calorimetry the inelastic cross sections, $\sigma_{\text{inel}}(\text{pA})$ or $\sigma_{\text{inel}}(\pi\text{A})$, determine the value of the corresponding interaction length, $\lambda_{\text{int}} = A/N_{\text{A}}\sigma_{\text{inel}}(\text{hadron, A})$. On geometrical grounds $\sigma_{\text{inel}}(\text{hadron, A})$ is expected to scale as $A^{2/3}\sigma_{\text{inel}}(\text{hadron, p})$, close to the measured approximate scaling $A^{0.71}\sigma_{\text{inel}}(\text{hadron, p})$ and therefore $\lambda_{\text{int}} \approx A^{0.29}/\{N_{\text{A}}\sigma_{\text{inel}}(\text{hadron, p})\}$ [g cm^{-2}].

This characteristic length λ_{int} is the mean free path of high energy hadrons between hadronic collisions and sets the scale for the longitudinal hadronic shower profile. The probability $P(z)$ for a hadron traversing a distance z without undergoing an interaction is therefore $P(z) = \text{exp.}\ (-z/\lambda_{\text{int}})$. The equivalence with the characteristic distance X_0 for the electromagnetic cascade is evident. In analogy to the parameterization of electromagnetic showers the longitudinal profile of hadronic showers can be parameterized in the form.

$$\mathrm{d}E/\mathrm{d}x = c\left\{w\{x/X_0\}^{\alpha-1}\exp\left(-bx/X_0\right) + (1-w)\,(x/\lambda)^{\alpha-1}\exp\left(-dx/\lambda\right)\right\}. \tag{6.24}$$

The overall normalization is given by c; α, b, d, w are free parameters and x denotes the distance from the shower origin [36].

Longitudinal pion-induced shower profiles are shown in Fig. 6.19 for different energies together with the analytical shower fits. The longitudinal energy deposit rises to a maximum, followed by a slow decrease due to the predominantly low-energy, neutron-rich part of the cascade. Proton-induced showers show a slightly different longitudinal shape due to the differences in the first few initial collisions. Shower profiles in different materials, when expressed as a function of λ_{int} exhibit approximate scaling in λ_{int}, in analogy to approximate scaling of electromagnetic

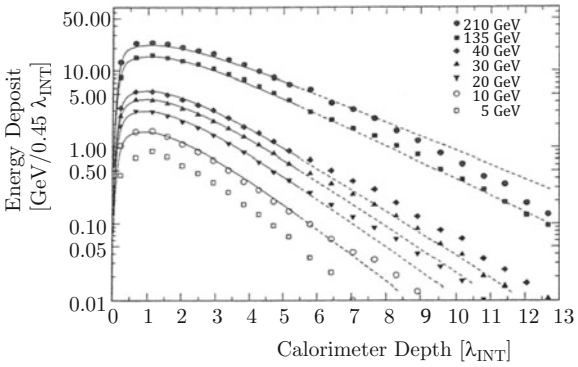


Fig. 6.19 Measured longitudinal shower distributions for pions at three energies together with the shower parameterization [37]

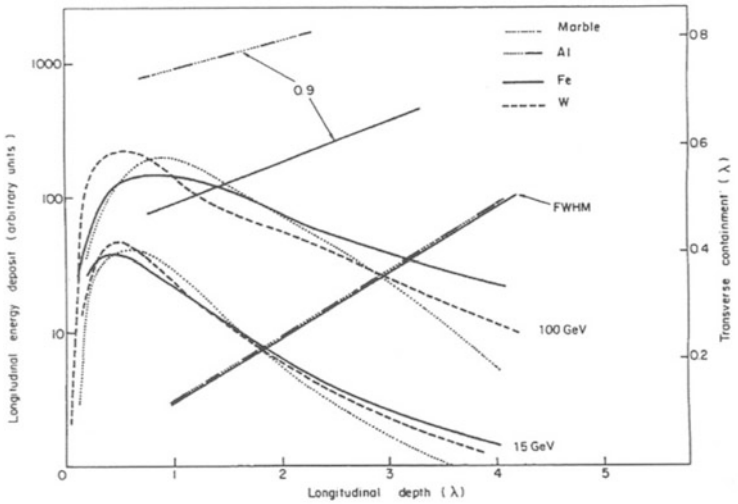


Fig. 6.20 Longitudinal shower development induced by hadrons in different materials, showing approximate scaling in λ. The shower distributions are measured with respect to the face of the calorimeter (left ordinate). The transverse distributions as a function of shower depth show scaling in λ for the narrow core. The 90% containment radius is much larger and does not scale with λ (right ordinate) [30]

showers in X_0, see Fig. 6.20. Also shown are the transverse shower distributions: the relatively narrow core is dominated by the high-energy (mostly electromagnetic) component. The tails in the radial distributions are due to the soft, neutron-rich, component. In Fig. 6.21 the fractional containment as a function of energy is shown, exhibiting approximately the expected logarithmic energy dependence for a given containment [38, 39].

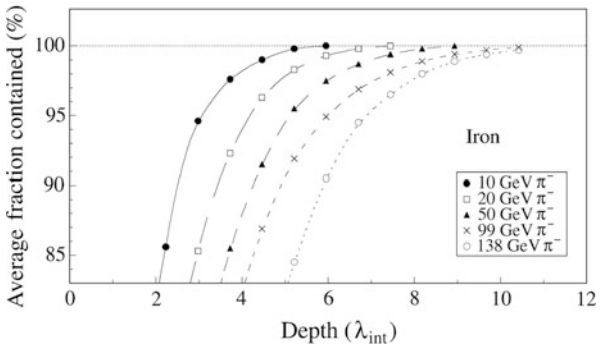


Fig. 6.21 Measured average fractional containment in iron of infinite transverse dimension as a function of thickness and various pion energies [38, 39]

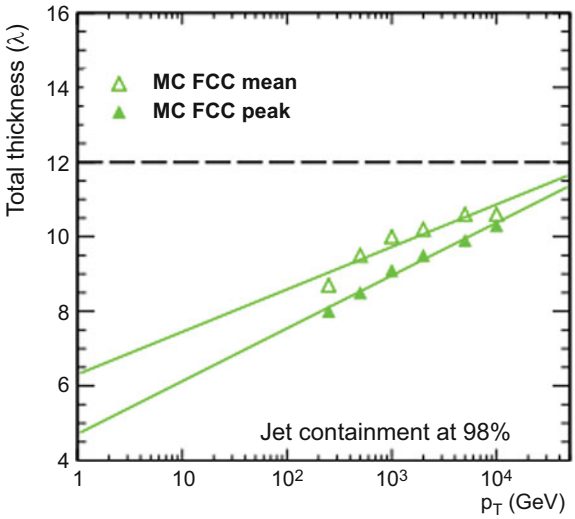


Fig. 6.22 Total thickness, expressed in λ, to contain up to 98% of a jet as a function of the jet transverse momentum. Mean and peak refer to different statistical measures of containment [40]

These results indicate that for 98% containment at the 100 GeV scale a calorimeter depth of 9 λ_{int} is required. At the LHC, where single particles energies in the multi-hundred GeV and jets in the multi-TeV range have to be well measured, the hadrons are typically measured in 10 λ_{int}. For the next jump in collider energy, as is presently studied e.g. for "Future Circular Collider, FCC", particle and jet energies are approximately a factor 10 higher. For adequate containment, i.e. at the 98% level, calorimeter systems with ~12 λ_{int} will be required, see Fig. 6.22 [40]

6.2.7 Energy Resolution of Hadron Calorimeters

The average properties of the hadronic cascade are a reflection of the intrinsic event-by-event fluctuations which determine the energy resolution. Most importantly, fluctuations in the hadronic component are correlated with the number of spallation neutrons and (delayed) nuclear photons and hence with the energy consumed to overcome the binding energy; these particles from the nuclear reactions will contribute differently (in general less) to the measurable signal.

Let η_e be the efficiency for observing a signal $E^e{}_{vis}$ (visible energy) from an electromagnetic shower, i.e., $E^e{}_{vis} = \eta_e\ E(em)$; let η_h be the corresponding efficiency for purely hadronic energy to give a measurable signal in an instrument. Decomposing a hadron-induced shower into the em fraction F_{em} and a purely hadronic part F_h the measured, 'visible' energy $E^\pi{}_{vis}$ for a pion-induced shower is.

$$E^\pi{}_{vis} = \eta_e F_{em} E + \eta_h F_h E = \eta_e (F_{em} + \eta_h/\eta_e F_h)\ E, \tag{6.25}$$

where E is the incident pion energy. The ratio of observable signals induced by electromagnetic and hadronic showers, usually denoted 'e/π', is therefore

$$E^\pi{}_{vis}/E^e{}_{vis} = (e/\pi)^{-1} = F_{em} + \eta_h/\eta_e F_h = 1 + (\eta_h/\eta_e - 1)\ F_h. \tag{6.26}$$

In general $\eta_e \neq \eta_h$: in this case, the average response of a hadron calorimeter as a function of energy will not be linear because F_h decreases with incident energy. More subtly, for $\eta_h \neq \eta_e$, event-by-event fluctuations in the F_h and F_{em} components produce event-by-event signal fluctuations and impact the energy resolution of such instruments. The relative response 'e/π 'turns out to be the most important yardstick for gauging the performance of a hadronic calorimeter.

A convenient (albeit non-trivial) reference scale for the calorimeter response is the signal from minimum-ionizing particles (*mip*) which in practice might be an energetic through going muon, rescaled to the energy loss of a mip. Let *e/mip* be the signal produced by an electron relative to a mip. Assume the case of a mip depositing e.g. α GeV in a given calorimeter. If an electron depositing β GeV produces a signal β/α, the instrument is characterized by a ratio $e/mip = 1$. Similarly, the relative response to the purely hadronic component of the hadron shower is $\eta_h F_h E/mip$, or *h/mip* which can be decomposed into $h/mip = (f_{ion}\ ion/mip + f_n\ n/mip + f_\gamma\ \gamma/mip)$, with f_{ion}, f_n, f_γ denoting the average fractions of ionizing particles, neutrons and nuclear photons.

Practical hadron calorimeters are usually built as sampling devices; the energy sampled in the active layers, f_S (Eq. 6.20), is typically a small fraction, a few percent or less, of the total incident energy. The energetic hadrons lose relatively little energy ($\leq$10%) through ionization before being degraded to such low energies that nuclear processes dominate. Therefore, the response of the calorimeter will be

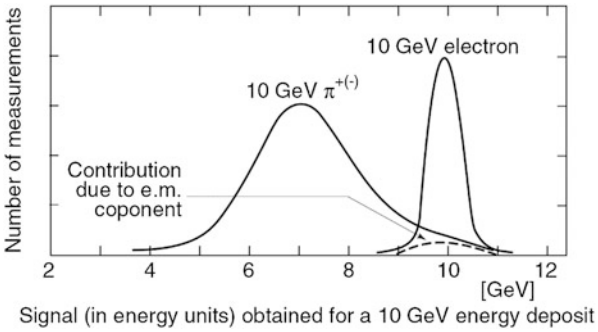


Fig. 6.23 Conceptual response of a calorimeter to electrons and hadrons. The curves are for a 'typical' sampling calorimeter with electromagnetic resolution of $\sigma/E = 0.1/\sqrt{E}(\text{GeV})$, with hadronic resolution of $\sigma/E = 0.5/\sqrt{E}(\text{GeV})$ and $e/\pi = 1.4$. The hadron-induced cascade fluctuates between almost completely electro-magnetic and almost completely hadronic energy deposit, broadening the response and producing non-Gaussian tails

strongly influenced by the values of *n/mip* and γ*/mip* in both the absorber and the readout materials.

This simple analysis already provides the following qualitative conclusions for instruments with $e/\pi \neq 1$, as shown conceptually in Fig. 6.23:

- fluctuations in $F_{\pi 0}$ are a major contribution to the energy resolution;
- the average value (F_{em}) increases with energy: such calorimeters have a non-linear energy response to hadrons;
- these fluctuations are non-Gaussian and therefore the energy resolution scales weaker than $1/\sqrt{E}$.

This understanding of the impact of shower fluctuations suggests to 'tune' the e/π response of a calorimeter in the quest for achieving e/π = 1, and thus optimizing the performance [41, 42].

It is instructive to analyze n/mip, because of the richness and intricacies of n-induced nuclear reactions and the very large number of neutrons with $E_n < 20$ MeV. In addition to elastic scattering a variety of processes take place in high-Z materials such as (n, n'), (n, 2n), (n, 3n), (n, fission). The ultimate fate of neutrons with energies $E_n < 1$–2 MeV is dominated by elastic scattering; cross-sections are large (~ barns) and mean free paths short (a few centimetres); the energy loss is ~1/A (target) and hence small. Once thermalized, a neutron will be captured, accompanied by γ-emission.

This abundance of neutrons gives a privileged role to hydrogen, which may be present in the readout material. In an n-p scatter, on average, half of the neutron kinetic energy is transferred. The recoil proton, if produced in the active material, contributes directly to the calorimeter signal, i.e., is not sampled like a mip (a 1 MeV proton has a range of ~20 μm in scintillator). The second important n-reaction is the production of excitation photons through the (n,n',γ) reaction [42].

This difference in neutron response between high-Z absorbers and hydrogen-containing readout materials has an important consequence. Consider the contributions of *n/mip* as a function of the sampling fraction f_S. The mip signal will be inversely proportional to the thickness of the absorber plates, whereas the signal from proton recoils will not be affected by changing f_S: the *n/mip* signal will increase with decreasing f_S. Changing the sampling fraction allows to alter, to 'tune' e/π. Tuning of the ratio R_d = passive material [mm]/active material [mm] is a powerful tool for acting on e/π [41]. This approach works well for high-Z absorbers with a relatively large fission cross section, accompanied by multiple neutron emission. Optimized ratios tend to imply for practical scintillator thicknesses rather thick absorbers with concomitant significant sampling fluctuations and reduced signals.

How tightly are the various *fluctuating* contributions to the invisible energy correlated with the *average* behaviour, as measured by e/π? A quantitative answer needs rather complete shower and signal simulations and confirmation by measurement. Two examples are shown in Fig. 6.24. One observes a significant reduction in the fluctuations and an intrinsic hadronic energy resolution of $\sigma/E \approx 0.2/\sqrt{E(\text{GeV})}$ for instruments with $e/\pi \approx 1$ [39, 41, 42]. The intrinsic

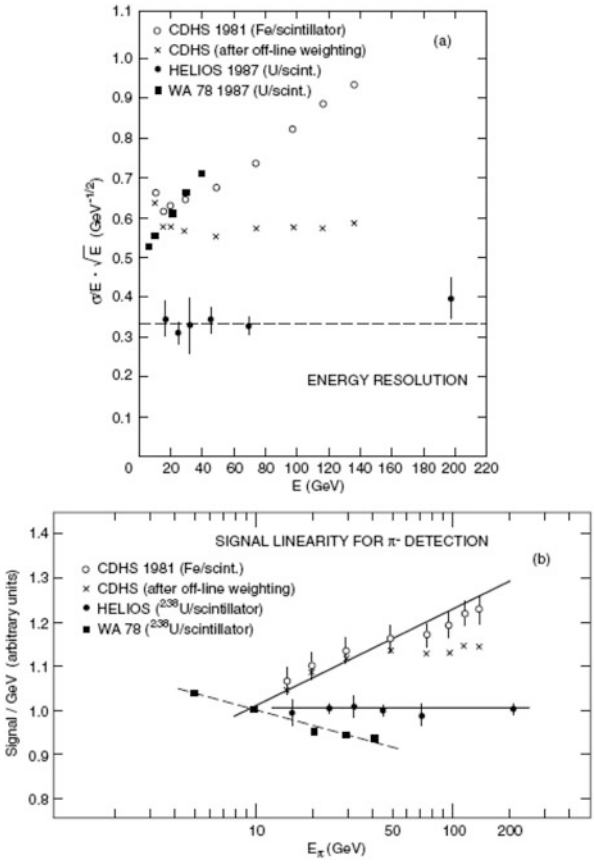


Fig. 6.24 Experimental observation of the consequences of $e/\pi \neq 1$. Shown is the measured pion response in under-compensating, compensating and over-compensating calorimeters; (**a**) energy resolution $\sigma/E\ \sqrt{E}$ as a function of the pion energy, showing deviations from scaling for non-compensating devices. (**b**) Signal per GeV as a function of pion energy, exhibiting signal non-linearity for non-compensating detectors [41]

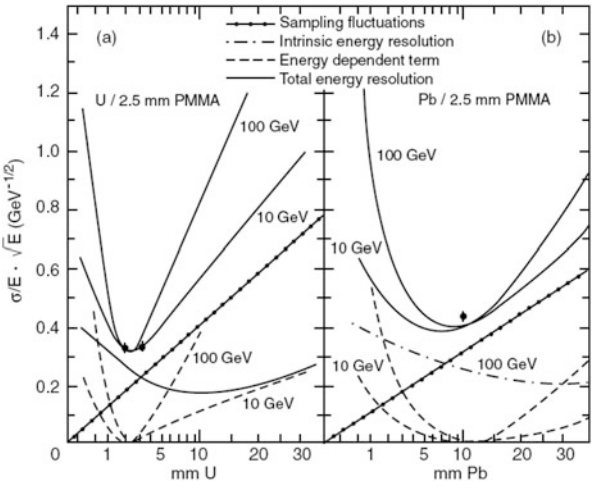


Fig. 6.25 Contributions to and total energy resolution of 10 and 100 GeV hadrons in scintillator calorimeters as a function of thickness of (**a**) uranium plates and (**b**) lead plates. The scintillator thickness is 2.5 mm in both cases. The dots in the curves are measured resolution values of actual calorimeters [42]

hadron resolution of a lead-scintillator sampling calorimeter may even be as good as $\sigma/E < \approx 0.13/\sqrt{E}(\text{GeV})$ [43].

Detectors achieving compensation for the loss of non-detectable ('invisible') energy, i.e., $e/\pi = 1$, are called 'compensated' calorimeters.

There are several further negative consequences if $e/\pi \neq 1$ in addition to reduced resolution. The energy resolution which no longer scales with $1/\sqrt{E}$, is usually parameterized as $\sigma/E = a_1/\sqrt{\text{E}} \oplus a_2$, where a 'constant' term a_2 is added quadratically, even though physics arguments suggest $a_2 = a_2(E)$. Since the fraction of π^0-production $F_{\pi 0}$ increases with energy, such calorimeters have a non-linear energy response. Furthermore, given that the average hadronic fraction F_h are different for pions ($F_\text{h}(\pi)$) and protons (neutrons) ($F_\text{h}(\text{p})$), typically $F_\text{h}(\pi) \sim 0.85 F_\text{h}(\text{p})$, the response in calorimeters with $e/\pi \neq 1$ depends on the hadron species [42].

The effects of e/π have been observed [41] (Fig. 6.24) and evaluated quantitatively [42]. Measurements and Monte Carlo simulations of the response of various calorimeter configurations are shown in Figs. 6.25 and 6.26.

Besides achieving "intrinsic compensation" with $e/\pi = 1$, effective compensation can be achieved by recognizing event by event independently the em fraction F_em and the hadronic fraction F_h, respectively. In instruments with a fine-grained longitudinal and lateral subdivision the different em and hadronic shower shapes provide an approximately independent determination of the two components and the basis for their off-line weighting, resulting in an effective $e/\pi = 1$ (see Sect. 6.7.5). Alternatively, the em component and the hadronic component in the shower may be measured independently with a dual readout: one active medium is only sensitive to Cherenkov radiation, predominantly caused by the em component, while the charged particles are measured e.g. with a scintillator, see Sect. 6.3.3.

To complete the analysis of the contributions to the energy resolution we need to consider sampling fluctuations, assuming fully contained showers and no degradation due to energy leakage. For electro-magnetic calorimeters a simple expla-

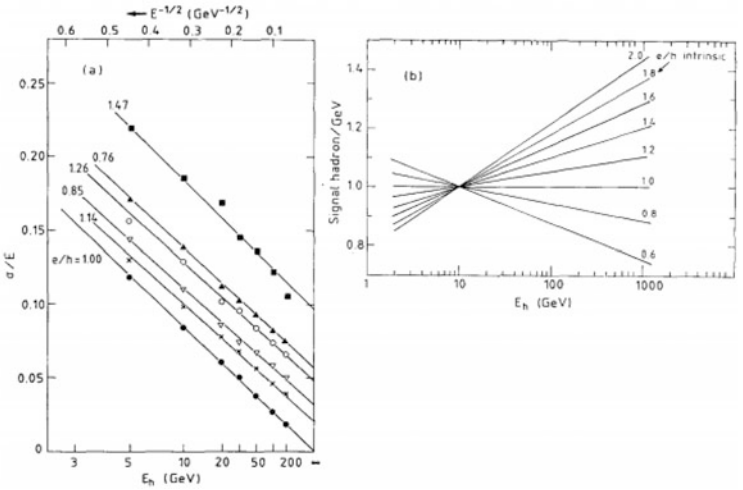


Fig. 6.26 Monte Carlo simulation of the effects of $e/\pi \neq 1$ on energy resolution (**a**) and linearity (**b**) of hadron calorimeters [42]

nation and an empirical parameterization holds (Eq. 6.21): $\sigma_{\text{samp}}(\text{em})/E = c(\text{em}) \cdot (\Delta E(\text{MeV})/E(\text{GeV}))^{1/2}$, where ΔE is the energy lost in one sampling cell and $c(\text{em}) \approx 0.05$ to 0.06 for typical absorber and readout combinations.

Similar arguments apply for the hadronic cascade; empirically, one has observed [30, 43] that.

$$\sigma_{\text{samp}}\,(\text{h})\,/E = c\,(\text{h}) \cdot (\Delta E\,(\text{MeV})\,/E\,(\text{GeV}))^{1/2} \text{ with } c\,(\text{h}) \approx 0.10. \tag{6.27}$$

For high-performance hadron calorimetry sampling fluctuations cannot be neglected.

The foundations of modern, optimized hadron calorimetry can be summarized as follows:

- the key performance parameter is $e/\pi = 1$, which guarantees linearity, $E^{-1/2}$ scaling of the energy resolution, and best intrinsic resolution;
- by proper choice of type and thickness of active and passive materials the response can be tuned to obtain (or approach) $e/\pi \sim 1$;
- the intrinsic resolution in practical hadron calorimeters can be as good as $(\sigma/E) \cdot \sqrt{E} < \sim 0.2$;
- sampling fluctuations contribute at the level of $\sigma/E \approx 0.10\ (\Delta E(\text{MeV})/E(\text{GeV}))^{1/2}$.

6.2.8 *Muons in a Dense Material*

The velocity dependence of the average energy loss by collisions of singly charged particles (muons, pions, protons, ...) with electrons of the traversed medium differs slightly from formula (6.1) and is given by:

$$-\frac{dE}{dx} = k\frac{Z}{A}\frac{1}{\beta^2}\left[\ln\frac{2m_e c^2\gamma^2\beta^2}{I} - \beta^2 - \frac{\delta}{2}\right] \left(\mathrm{MeV}/\left(\mathrm{g/cm^{-2}}\right)\right) \tag{6.28}$$

where $\delta \approx \ln(\gamma)$ accounts for screening effects at high energy. As a function of energy of the incident particle the most probable value shows a slow increase (relativistic rise) followed by a plateau whose value depends on the density of the material. The energy loss reaches a minimum for $\gamma\beta \sim 3$, corresponding to muon energies of few hundred MeV.

At a given energy, the energy loss distribution of $-\mathrm{d}E/\mathrm{d}x$ in a slab of material has an asymmetric distribution around its most probable value, usually referred to as the "Landau-Vavilov" distribution [44, 45]. The muon energy loss in dense materials has been extensively studied [46]. Both, the absolute energy loss and the straggling function agree with measurements at the percent level [47] up to several hundred GeV.

For muon energies above ~100 GeV, bremsstrahlung, pair production and deep inelastic scattering start to contribute, generating tails in the energy distribution ('catastrophic energy loss') [48, 49]. As an illustration, the average contribution of these processes for muons in iron up to 100 TeV is shown in Fig. 6.27. Very roughly speaking a muon behaves as an electron with a critical energy scaled as $\approx (m_\mu/m_e)^2$. However, unlike for electrons or positrons, pair production is larger than bremsstrahlung.

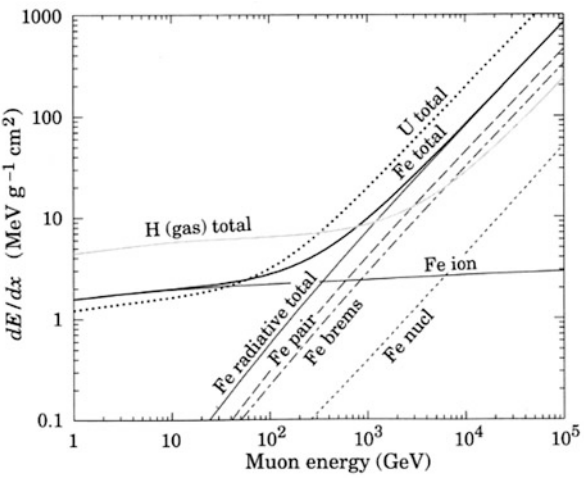


Fig. 6.27 Contributions to the energy loss of muons in iron, as a function of the muon incident energy. The total energy loss in hydrogen gas and uranium is also shown

Momentum correction to muon momenta can be applied, in setups where muons traverse a calorimeter before entering the muon spectrometer. For muons above ~10 GeV/c there is a good correlation between the total energy loss of muons in a calorimeter with the energy loss recorded in the active medium.

This is valuable, particularly for 'catastrophic' muon energy loss. Event-by-event correction for the muon energy loss is therefore useful in the hundred GeV momentum range for muon spectrometers behind the calorimeter with few percent momentum resolution [39].

Energy calibration and monitoring is frequently and conveniently done with muons. Exposing a calorimeter to a beam of electrons with well-known energy sets the 'electron-energy scale'.

In sampling calorimeters muons deposing a given energy produce in general more signal than electrons having deposited the same energy: $e/\mu < 1$. While establishing an absolute energy scale with muons requires very careful MC cross-checks, it is very convenient to use muons as a monitor of the calorimeter response as a function of time during data taking and as intercalibration tool between different parts of a calorimeter set-up [50]. The use of muons allows to transfer the absolute energy calibration established in a test beam to the experimental facility and to follow the energy calibration in situ using muons from physics channels. However, given the large dynamic range of energy measurements in many experiments, e.g. at the LHC and the smallness of the muon signal, complimentary calibration methods are necessary to achieve the required accuracy, see Sect. 6.3.6.

6.2.9 Monte Carlo Simulation of Calorimeter Response

Modern calorimetry would not have been possible without extensive shower simulations.

The first significant use of such techniques aimed to understand electromagnetic calorimeters. For example, electromagnetic codes were used in the optimization of NaI detectors in the pioneering work of Hofstädter, Hughes and collaborators [51]. One code, EGS4, has become the de facto standard for electromagnetic shower simulation [17]. Early hadronic cascade simulations were motivated by experimental work in cosmic-ray physics [52] and sampling calorimetry [53]. However, it were the codes developed by the Oak Ridge group [54], with their extensive modelling of nuclear physics, neutron transport, spallation and fission, which are indissociable from the development of modern hadron calorimetry [35].

Modern, high precision calorimetry and related applications have imposed a new level of stringent quality requirements on simulation:

- in many applications, electromagnetic effects have to be understood at the 0.1% level, hadronic effects at the 1% level;
- 'unorthodox' calorimeter geometries (Sect. 6.7) have to be optimized with simulation tools providing sophisticated interfaces to shower codes:

– in modern calorimeter facilities the energy deposits are usually distributed over several systems of different geometries and materials. Simulation codes are pushed to their limits in translating the recorded signal into a 1% precision energy measurement;
– at LHC and in particular in the study of the UHE Cosmic Ray Frontier simulation codes are used to extrapolate measured detector response by one to eight (!) orders of magnitude;
– particle physics MC codes are applied to areas outside particle physics, such as of radiation shielding, nuclear waste incineration and medical radiation treatment.

First, we will describe the general approach to these simulation issues before addressing some specific points. Regular conferences on this subject provide a good overview [55].

Electromagnetic Shower Simulation
For decades EGS4 [17] has been the standard to simulate electromagnetic phenomena. A modern extended incarnation has been developed by the GEANT4 Collaboration [18]. It includes the full panoply of radiation effects, including photons from scintillation, Cherenkov and Transition radiation up to electromagnetic phenomena relevant at 10 PeV.

Hadronic Shower Simulation
The simulation must cover the physics and the corresponding cross-sections from thermal energies (neutrons) up to (in principle) the 10^{20} eV frontier, requiring many different physics models; program suites, 'toolkits', such as GEANT4 [18], provide the user with choices of physics interaction models to select the physics interactions and particle types appropriate to a given experimental situation.

At high energies (~15 GeV to ~100 TeV)—in addition to measured cross sections—models describing the hadron physics are used, such as the 'Quark Gluon String' model [18], Fritiof or Dual Parton Models [56]. Such models are coupled to descriptions of the fragmentation and de-excitation of the damaged nucleus. At the highest energies other models, such as 'relativistic Quark Molecular Dynamics' models are being developed [57].

In the intermediate energy range (<10 GeV) Bertini-style cascade models [58] are employed to describe the intra-nuclear cascade phenomena. These models use measured cross-sections and angular distributions.

For the very low energy (<20 MeV) domain neutron transport codes have been developed, using experimental cross-sections.

The different energy regimes covered by these models are connected with parametric descriptions, in which cross-sections are parameterized and extrapolated over the full range of hadronic shower energies. Well-known examples are Geisha [59] and to a certain extent GCalor (or GEANTCalor) [60].

Applications: Illustrative Examples
We present comparisons of simulation with experiment to illustrate the quality of shower modelling.

(i) Energy Calibration and Reconstruction

Many physics programmes at the modern colliders (HERA, Fermilab, LHC) require energy measurements at the limit of the instrumental resolution and with ~1% accuracy. The calorimeters are frequently composed of different electromagnetic and hadronic instruments, made from different materials and sampling topologies.

Establishing the absolute energy scale in the reconstruction of particles (and jets) needs a major effort to understand the detector, from an instrumental and technical point. It requires a tight interplay between measurements and simulations. Energy calibration and reconstruction, proceeds in several steps. Customarily, a calorimeter (segment) is exposed to electrons, setting the 'electromagnetic' energy scale. For hadrons a 'weighing' has to be applied to each cell, such that.

$$E_{\mathrm{i}}\,(\mathrm{true}) = w_{\mathrm{i}} E_{\mathrm{i}}\,(\mathrm{reconstructed}) \;\text{ with } w_{\mathrm{i}} = \langle E_{\mathrm{i}}\,(\mathrm{true})\,/E_{\mathrm{i}}\,(\mathrm{reconstructed})\rangle\,.$$

E_{i}(true) expresses the total energy deposited. This can be a rather large correction, particularly in non-compensating calorimeters. In a further step, details of the energy reconstruction algorithm ('clustering') are simulated to evaluate the energy outside the cluster, usually chosen smaller than the true shower extent. In practical calorimeters, non-sensitive regions ('dead material', DM) are unavoidable leading to frequently sizeable corrections evaluated by MC.

Establishing the energy scale for jets is the most complex calibration task. Jets are calibrated with a series of simulation-based corrections and in situ techniques. In situ techniques exploit the transverse momentum balance between a jet and a reference object such as a photon, Z boson or multijet system for jets with $20 < p_T < 2000$ GeV, using both data and simulation. In this way an uncertainty in the jet energy scale approaching 1% is obtained for high-p_T-jets with $100 < p_T < 500$ GeV/ c. An uncertainty of about 4.5% is found for low-p_T jets ($p_T < 20$ GeV/ c), dominated by uncertainties in the corrections for multiple proton-proton interactions (pile-up), see Fig. 6.28 [61].

(ii) Particle Flow Analysis in Calorimeter Systems at Present and Future Colliders

An important recent development is an ambitious analysis strategy for reconstructing the jet energy in calorimeters, the "Particle Flow" concept. It aims at identifying and reconstructing individually each particle arising from the collision (proton-proton, electron-positron, ...) by combining the information from all the subdetectors. The resulting particle-flow event reconstruction leads to an improved performance for the reconstruction of jets and "Missing Transverse Energy" (MET). The algorithm also improves the identification of electrons, muons, and taus. While the concept has first been applied in the physics analysis at the LEP collider, it is presently heavly used by the LHC collaborations [62, 63]. The improvement can be dramatic, as shown in Fig. 6.29.

The benchmark performance for calorimeter systems (Sect. 6.7.6.2) for future colliders (International Linear Collider, ILC; Future Circular Collider, FCC) aims at a jet energy resolution of $\sigma(\mathrm{jet}) \sim 0.3/\sqrt{E(\mathrm{GeV})}$. This is motivated by the need to measure, e.g. W- and Z-decays into two jets with a mass resolution approaching

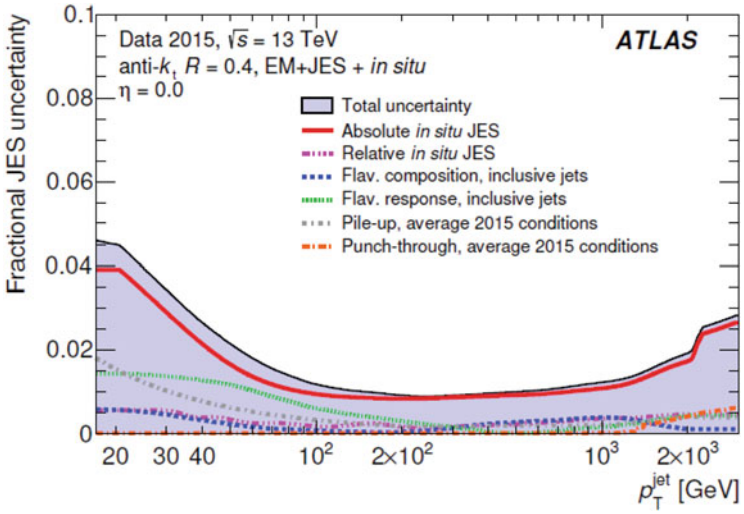


Fig. 6.28 Combined uncertainty in the jet energy scale (JES) of fully calibrated jets as a function of jet p_T in the central region of the ATLAS calorimeter system [61]

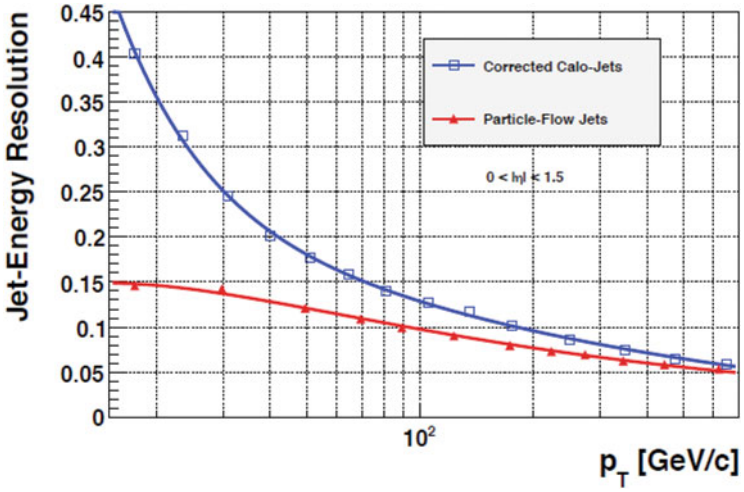


Fig. 6.29 Jet resolution for di-jets events in the CMS calorimeter reconstructed with the particle flow (red triangles) and the calorimeters (blue open squares) [63]

their natural width, i.e. with ~2 GeV (FWHM). Given that these jets are composed on average of ~60% hadrons, ~30% photons (the rest being shared by slow neutrons, neutrinos, muons, ...) a rather conventional resolution of $\sigma(\text{em}) \sim 0.15/\sqrt{E(\text{GeV})}$ and $\sigma(\text{hadronic}) \sim 0.5/\sqrt{E(\text{GeV})}$ would suffice, provided the individual energy deposits can be correctly associated with the individual particles measured in the charge particle spectrometer. This places a new level of performance requirements

on the calorimetry in terms of granularity, but also on the correct association of photonic and hadronic energy. Modeling has shown that this performance can be achieved in principle using the concept of 'Particle Flow Analysis'.[64, 65].

(iii) Ultra-High Energy Modelling
A particularly challenging application of these Monte Carlo techniques is extrapolation beyond present accelerator energies. The use of the Earth's atmosphere as a hadron calorimeter allows cosmic hadrons and nuclei up to and beyond 10^{20} eV to be probed. This requires 'dead-reckoning' of the detector response based on Monte Carlo techniques. Considerable faith in the extrapolation of the simulation models is needed in establishing the absolute energy scale. The estimate of the primary energy is based on measuring the shower shape: knowledge of F_{em}, the nucleon–nucleon cross-section, particle multiplicities, transverse momentum distributions, etc., all contribute to the estimate of the primary energy.

(iv) Low Energy Performance and Radiation Background
In many applications, e.g. dosimetry, careful modelling of the physics down to the MeV scale is needed. Certain codes [66] have been carefully benchmarked showing agreement to better than 20%, remarkable, as the very low-energy modelling of nuclear physics processes is involved.

Faithful modelling is also necessary to estimate the radiation levels in the LHC experimental caverns. Such modelling [67], based on the FLUKA code, was the basis for a number of design criteria and choices for the ATLAS and CMS experiments.

(v) Medical Applications
In cancer treatment with particle beams the tumour is exposed to proton or light ion beams, such as He or C^{12}, with energies of a few hundred MeV/nucleon. The energy deposition of the beam inside the human body (here the $1/\beta^2$ part of $\mathrm{d}E/\mathrm{d}x$ is relevant) can be monitored by positron emission tomography (PET), the β^+ emitters being produced through nuclear fragmentation reactions of the beam ions with the tissue nuclei.

Both, the patient treatment plan and the interpretation of these images is evaluated with the same MC programs as used in particle physics. More generally, the improvement in radiation treatments achieved with proper (particle physics) quality simulation is very significant, a very important legacy of particle physics to society [68].

We conclude that

- modern calorimetry owes much to Monte Carlo modelling;
- as always, predictions have to be taken with circumspection, in particular the extrapolation to performance and energy regimes inaccessible to experimental checks. Caveat emptor.

6.3 Readout Methods in Calorimeters

6.3.1 Scintillation Light Collection and Conversion

Scintillator materials used in calorimetry are inorganic crystals, organic compounds and noble liquids. Dense inorganic crystals represent one of the best techniques for homogeneous electromagnetic calorimetry. These crystals are insulators with a normally empty conduction band. When energy is deposited in the crystal, an electron can jump into the conduction band and cascade to the valence band by intermediate acceptor levels, part of the energy being emitted as light. The emitted light needs to be in the wavelength range where good photodetectors are available, and the crystal must be transparent to this wavelength range. The lifetime of the light emission depends on the concentration of acceptor levels, and temperature. In general, different decay times are present in the light luminescence spectrum of a given crystal (see also Chap. 3).

A list of commonly used scintillators, with some of their characteristic properties is given in Table 6.2. Crystals for homogeneous calorimetry are usually shaped as bars, typically of ~25 X_0 length and ~ 1 × 1 ρ_M transverse size. In colliding beam detectors, the cylindrical geometry leads in general to the use of tapered bars, with the incident radiation impinging on the smaller face. The growth of good quality ingots, followed by sawing and polishing to the needed size and surface quality requires specialized tooling available in industry. Careful packaging of the crystal in appropriate material (Tywek or equivalent) and sometimes lateral masking are needed to minimize the response dependence on position, transversally and longitudinally. The light detector (photomultiplier, photodiode, ...) is optically coupled to the back face of the crystal. The overall light yield, including the area and quantum efficiency of the transducer, influences the achievable energy resolution. A light yield of 1 photoelectron per MeV implies that the energy resolution cannot be better than $\sigma(E)/E = 3\%/\sqrt{E}$ (GeV). The number of emitted photons per MeV is in general much larger, being for example $4{\cdot}10^4$ in NaI doped with Thallium, one of the best scintillating crystal in terms of light yield. $PbWO_4$ produces ~150 times less light than NaI, but is far superior in other aspects (density, radiation resistance). New (and expensive) materials, like LYSO (a compound of Lutetium) are being developed for applications requiring fast response and high light yield.

A photomultiplier is schematically sketched in Fig. 6.30. All elements are located in an evacuated glass envelope. At the photocathode an electron is extracted by the photo-electric effect. A voltage difference accelerates the electron towards the first dynode out of which several electrons are extracted by secondary emission. This process is repeated over ~10 dynodes up to the anode at the highest (~1000 to 2000 volts) positive potential. With a sufficiently large gain at the first dynode the fluctuation of the number of electrons in the final charge pulse is dominated by the Poisson fluctuation of the number of photo-electrons. Amplification factors of several thousands are typical. A careful design of the High Voltage divider chain is mandatory to avoid non-linear effects. With recently developed "super bi-alkali"

Table 6.2 Properties of scintillating crystals applied in particle physics experiments

	NAI(Tl)	CsI(Tl)	CsI	BaF_2	CeF_3	BGO	$PbWO_4$	LYSO
Density [g cm^{-3}]	3.67	4.51	4.51	4.89	6.16	7.13	8.3	7.1
Radiation length [cm]	2.59	1.85	1.85	2.06	1.68	1.12	0.89	1.16
Molière radius [cm]	4.8	3.5	3.5	3.4	2.6	2.3	2.0	2.07
Interaction length [cm]	41.4	37.0	37.0	29.9	26.2	21.8	18.0	20.3
dE/dx)mip [MeV cm^{-1}]	4.79	5.61	5.61	6.37	8.0	8.92	9.4	9.2
Refractive index [at λ_{peak}]	1.85	1.79	1.95	1.50	1.62	2.15	2.2	1.8
Hygroscopicity	Yes	Slight	Slight	No	No	No	No	No
Emission spectrum, λ_{peak}								
Slow component [nm]	410	560	420	300	340	480	510	
Fast component [nm]			310	220	300		510	420
Light yield rel. to NaI								
Slow component	100	45	5.6	21	6.6	9	0.3	
Fast component			2.3	2.7	2.0		0.4	75
Decay time [ns]								
Slow component	230	1300	35	630	30	300	50	
Fast component			6	0.9	9		10	35

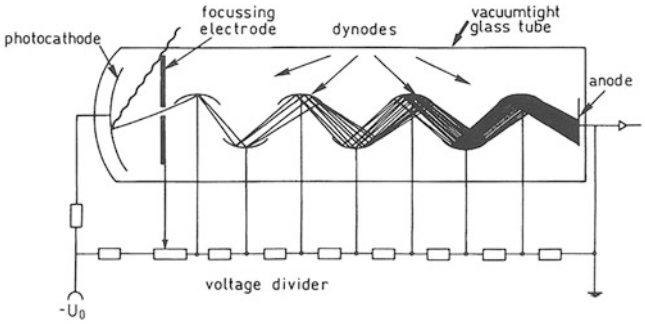


Fig. 6.30 Working principle of a photomultiplier. The electrode system is mounted in an evacuated glass tube

photocathodes (Cs-K) the quantum efficiency can reach more than 40% at 400 nm wavelength. For short wavelengths the efficiency is determined by the transparency of the entrance window. Quartz, CaF_2 or even LiF windows are necessary when efficiency in the near UV is required.

Because of their sensitivity to external magnetic fields, their rather large size and their cost, photomultipliers are nowadays being replaced by devices with less internal gain, followed by a high gain low-noise amplifier. Besides photo-triodes, the new devices are solid state based, like photodiodes or Avalanche Photo-Diodes (APD) [69]. Both offer good quantum efficiency, magnetic field insensitivity, moderate cost, small volume and—for APDs -a significant charge gain. The amplification is however accompanied by an "excess noise factor", of

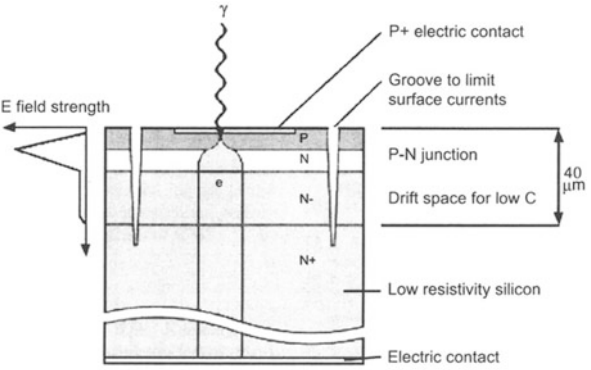


Fig. 6.31 Schematic diagram showing the structure of an avalanche photo-diode (APD)

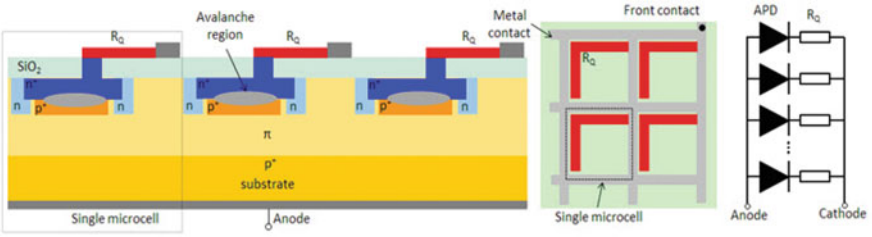


Fig. 6.32 Schematic diagram showing the structure of a Silicon photomultiplier (SiPM)

typically a factor 2 for a gain of ~50. This, together with the reduced size (and hence light collection) as compared to photocathodes can affect the energy resolution. The light detection and electron multiplication take place (see Fig. 6.31) in a thin layer (<40 μm) which lowers the sensitivity of APDs to minimum ionizing particles traversing the detector, as compared to simpler photodiodes.

The concept of APDs was extended to "Silicon Photomultipliers" by dividing the surface exposed to photons into small pixels, in a number large enough that each of them receives at most one photon.

Operating the device in the Geiger mode-i.e. with a very large gain-, and summing the current of a large number of pixels, one obtains effectively the equivalent of an analogue response to the number of incident photons, while each pixel operates in a binary mode.

Since the pioneering work [70], these devices have seen an extremely fast development [71]. A sketch of the layout of a SiPM is shown in Fig. 6.32.

Crystal calorimeters are the choice technology for precision electromagnetic calorimetry at medium energy machines like B-factories. CsI was used by Babar and Belle, and is used again for Belle II. The L3 experiment at LEP used BGO with success. However, the energy resolution reached for high energy electrons or photons (~50 GeV and above) was limited by the difficulty to calibrate a large

Table 6.3 Properties of noble liquids used in particle physics experiments

	LAr	LKr	LXe
Z	18	36	54
Boiling point [K]	87.3	119.8	165.0
Density in liquid phase [g cm^{-3}]	1.40	2.41	2.95
Radiation length [cm]	14.0	4.7	2.40
Molière radius [cm]	8.0	5.5	4.2
Nuclear interaction length for protons [cm]	84	61	57
Ionization properties			
Energy needed per electron-ion pair [eV]	24	17	15
Drift speed [mm/μs] at 10 kV/cm	5	3.8	2.6
Scintillation properties			
Emission spectrum, λ_{peak} [nm]	128	147	174
Decay time [ns]			
Fast component	5.0–6.3	2.0	2.2
Slow component	860–1090	80–91	27–34
Relative light yield in fast/slow component			
Fast component	8% (57%)	1%	5% (31%)
Slow component	92% (43%)	99%	95% (69%)
Refractive index at 170 nm	1.29	1.40	1.60

system (constant term of the energy resolution, see Eq. (6.23), of about 1% for the L3 BGO system) and not by the intrinsic resolution of the BGO crystals.

CMS and ALICE (for a part of its angular coverage) at the LHC decided to use $PbWO_4$. The most challenging case is CMS, given the very large size of the EM calorimeter, and the high radiation levels in the high luminosity collision points of the LHC, with nominally 500 fb $^{-1}$ of integrated luminosity at 14 TeV. More details are given in Sect. 6.7.3.

In some applications crystals are read on both ends, providing longitudinal information. However, so far it has not been possible to split the crystals longitudinally in independent segments without degrading the performances, a limitation for particle identification (see Sect. 6.4.3).

Noble liquids are also good, fast scintillators. Table 6.3 gives the properties of liquid argon, krypton and xenon already used in several practical cases for their scintillation properties.

In liquid argon about 4.10^4 photons are emitted per MeV deposited, a number very close to what is quoted for NaI. The light is however emitted in the far ultraviolet range, which complicates the conversion to electrical signals. Recent work [72] has shown that the scintillation light emitted by helium in the extreme vacuum ultraviolet range (~80 nm) can be used for particle detection, thanks to wavelength shifters (see below). The mechanism of scintillation in noble liquids involves the formation of excited diatomic molecules around the primary ions, which decay to free atoms by emitting radiation. In order to keep the emitted light associated with a well-defined region of space, thin reflecting boxes can be

introduced in the liquid volume. At present, one of the largest size detectors using light from noble liquids is the xenon calorimeter of the MEG experiment [73] (see also Sect. 6.7.1). As already mentioned in 6.2.3, the search for dark matter has triggered the development of several large size experiments using liquid xenon. These experiments [74] exploit both the scintillation and the ionization signal of the sought for nuclear recoils. Ionization electrons are preferentially transported to the surface of the liquid bath where, in a high electrical field region, they are extracted with high efficiency [74] and accelerated in the gas phase, giving in turn rise to (delayed) light emission. One example is described in Sect. 6.7.2.

Future long baseline neutrino experiments of very large size, like the DUNE [75] project at Fermilab envision liquid argon detectors of several tens of kilotons. DUNE will exploit both the scintillation and the ionization signals. In one of the read-out options, called "single-phase", the ionization signal is directly collected by a set of wires, each equipped with a readout chain, in order to have access to details of all secondary produced particles. The other option, "dual-phase", is close to what is described above for dark matter searches.

Liquid scintillators have been used abundantly in neutrino experiments, either in totally active large volume detectors, like Kamland and SNO, or as a large array of tubes filled with doped mineral oil.

The most recent example of the latter is NOvA [76] in which each tube is read out by means of a wavelength shifting fiber connected to a single pixel of an APD. The chapter on neutrino detectors provides further details.

Plastic scintillator plates, such as Polymethylmetacrylate (PMMA) doped with organic scintillator, have been used for electromagnetic and even more extensively for hadronic sampling calorimetry. The principal difficulty using this technology is the light extraction. The dimension of scintillator tiles of typically 10 cm $\times$ 10 cm size and 0.5 cm thickness would require light guides of typically 10 cm $\times$ 0.5 cm section in order to extract the light while preserving the emission phase space (respecting Liouville's theorem), a very difficult task in realistic detector layouts.

An elegant solution is the use of wavelength shifters [77, 78] in which due to their isotropic emission a constant fraction of the light is transported from the scintillating tile to a small rod, or even a plastic fibre separated from the tile by an air gap. The principle is shown in Fig. 6.33. Many calorimeter facilities at colliders were built following this principle, see also Sect. 6.7.

In a further development, detectors capable of accommodating smaller transverse granularities (like 5 cm $\times$ 5 cm) were proposed, like the "Shashlik" concept in which readout fibres cross the scintillating tile and the passive converter perpendicularly to their faces [79]. Originally considered in CMS, this scheme was later chosen by the LHCb experiment at the LHC for its electromagnetic calorimeter. A sketch of the arrangement of absorbers, scintillating tiles and fibers is shown in Fig. 6.34.

Even more ambitious was the "Spaghetti" calorimeter [80, 81] in which each calorimeter cell (typically 1 $\times$ 1 ρ_M transverse size and 25 X_0 deep) is built out of scintillating fibres embedded in a lead matrix, oriented parallel to the long side of the block. The electromagnetic calorimeter of the KLOE [82] experiment at the DAFNE

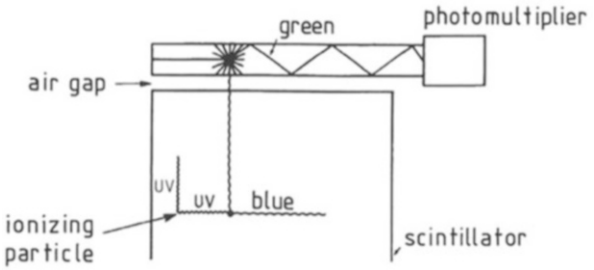


Fig. 6.33 Wavelength shifter readout of a scintillator

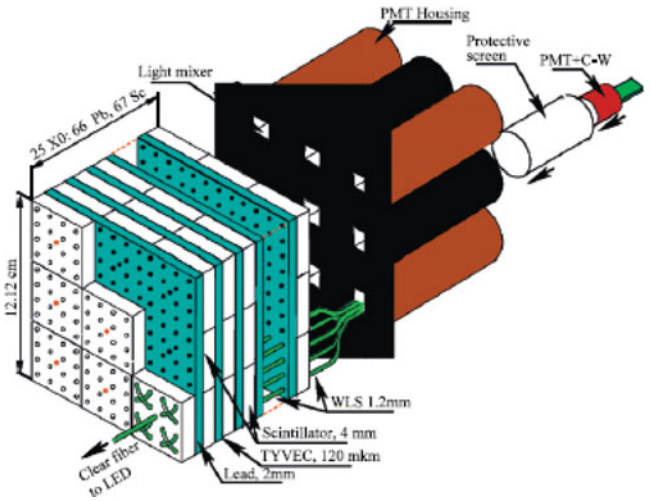


Fig. 6.34 The "shashlik" concept as realized in the LHCb Electromagnetic calorimeter

electron–positron collider in Frascati was built along these principles-although with a different geometry—and gave excellent results in the energy range of this machine.

6.3.2 Cherenkov Light Collection and Conversion

Although much less intense than scintillation light in good scintillators, Cherenkov radiation represents in some cases an interesting alternative. When a charged particle (electron or positron in the case of an electromagnetic shower) propagates in a transparent medium with a speed βc, larger than the speed of light c/n in this medium, an electromagnetic wave forms along a cone of half-angle $\theta_c = A\cos$

(1/βn) with respect to the incident particle direction, and with a number N of emitted photons in the visible range (400 to 700 nm) per unit length:

$$\mathrm{d}N/\mathrm{d}x = 490 \sin^2\theta_\mathrm{c} \left[\mathrm{cm}^{-1}\right]. \tag{6.29}$$

Lead glass, a dense material with a high index of refraction, has been used in several experiments (in particular OPAL [83] at LEP) with very similar geometries (tapered bars) as described above for scintillating crystals. The energy resolution is limited by the number of electrons and positrons in the shower above the Cherenkov threshold, resulting in a stochastic term $\sigma(E)/E$ of $> \sim$ 5–6%/$\sqrt{E}$, comparable to very good sampling calorimeters. Given the small number of photons, readout with photomultipliers is mandatory. As for crystals, longitudinal segmentation is in general not feasible. In several cases, "preshowers" of a few X_0 depth, instrumented with another higher granularity readout technique, have been used in front of lead glass arrays, in order to improve particle identification (see Sect. 6.4.3). Another limitation for large collider systems is the reduced response of lead glass to hadronic showers (a large fraction of the hadronic cascade is made of non-relativistic particles), inducing a performance limitation for hadronic calorimetry. However, the preponderance of Cherenkov-light production from electrons and positrons, i.e. the electromagnetic part of the hadronic shower, offers an interesting possibility. A hadronic sampling calorimeter instrumented with two sets of fibres—one set sensitive to Cherenkov-light only, the other set consisting of scintillating fibres, sensitive to all charged particles—can measure separately the electromagnetic component of the hadronic shower. This possibility is being studied in the dual-readout "DREAM" project. Test beam results are reported in Ref. [84].

Exploiting only the Cherenkov component, an hadronic calorimeter made of quartz fibers (parallel to the beam axis) embedded in an iron matrix has been chosen for the very forward calorimeter of the CMS experiment (for the pseudorapidity region up to 5). This choice was motivated by the high radiation resistance of quartz fibers, well adapted to this harsh environment [85].

Energy measurement with Cherenkov light produced in water was used with great success in very large detectors for nucleon decay and solar neutrino experiments, like Superkamiokande [86]. For the required detector volume of 50,000 tons water, the Cherenkov light was read out using large photomultipliers. In Superkamiokande, 50% of the outer surface of the detection volume is covered by 50 cm diameter phototubes. Electrons of 10 MeV are reconstructed with an energy resolution of about 15%. Their position in the detector volume is reconstructed with an accuracy of 70 cm and their direction with an accuracy of ~25 degrees. The detector also provides some discrimination between electrons (showering) and muons (single Cherenkov cone).

6.3.3 *From Ionization to Electrical Signal in Dense Materials*

One major avenue for calorimetry instrumentation is the measurement of the ionization charge produced in dense, active materials. In the presence of an applied electric field the charges move, inducing a current in readout electrodes proportional to the liberated charge and hence to the energy deposited by the showering particle. Electric charges are much easier to transport and to collect compared to light, which is the basic, decisive advantage of this concept.

This technique was introduced in the early 1970s [87] using liquefied argon as the active material. It has matured into one of the most widely used methods for calorimetry instrumentation, in particular, of sampling calorimeters. Noble liquid ionization calorimeters offer a number of attractive advantages, especially for instruments in the difficult environment of colliders. They are characterized by intrinsic stability and excellent uniformity of response (the only amplification is in the electronics chain which is fairly easy to calibrate), relative ease of a high segmentation and reasonable cost.

Other materials than argon are suitable for this method of detection, in particular the heavier noble liquids (Kr, Xe). In liquid helium and liquid neon, electrons are trapped in nano-scale cavities, and drift with characteristic speeds about a thousand times slower than electrons in other noble liquids. Solid neon was found to be usable at low rate [88]. Some saturated molecules like Tetramethylpentane (TMP), which is a liquid at room temperature, have also been tried. High purity at the ppb-level, required to avoid electron trapping, has limited their use compared to noble liquids, which however require cryogenic operation. The properties of noble liquids for ionization calorimetry are given in Table 6.3. Besides the value of $\mathrm{d}E/\mathrm{d}x$ and X_0 specific to the material, important parameters are the mean energy needed to create an electron-ion pair, the electron drift speed as a function of the electric field, and the dielectric constant, which affects the capacitance of a readout cell. Since the ions have a much smaller drift velocity compared to electrons, a track crossing a gap (and depositing charge uniformly) will give rise to a triangular current (see Fig. 6.35) given by Eq. (6.30) where $+Q_0$ and $-Q_0$ are the liberated charges, d the gap, and v the drift velocity of electrons. The resulting current is

$$I(t) = Qv/d \tag{6.30}$$

with $Q = Q_0(1 - vt/d)$. This formula is easily derived by remembering that a point charge q at a distance x from one of the parallel planar electrodes defining the gap of width d, induces a charge $-q(d - x)/d$ on this electrode, and $-xq/d$ on the other one.[2]

[2]In case of test cells with a grid at an intermediate potential in between the two electrodes, all charges of the grid-cathode region contribute with the same weight to the anode signal.

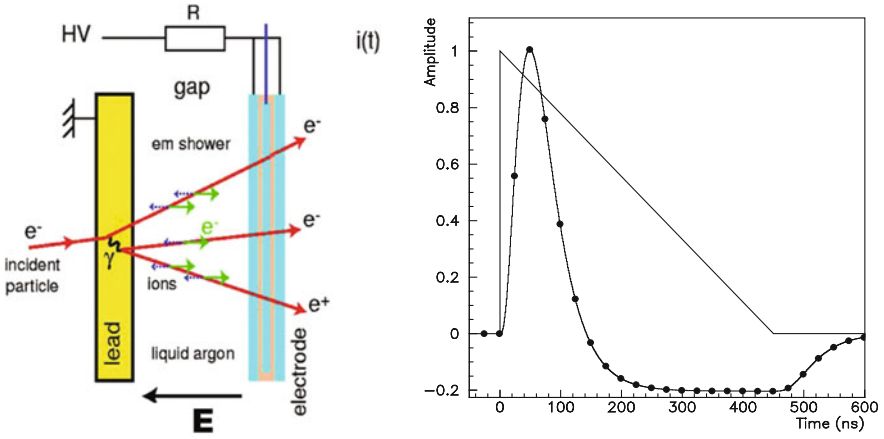


Fig. 6.35 Current induced by charges drifting in the sensitive gap of an ionization calorimeter. Left: charges drifting in the gap; right: current from drifting charges (triangle), and after CR-RC2 shaping. The dots every 25 ns represent times where the signal is being sampled (40 MHz sampling)

Depending on the rate of particles hitting a given cell, the readout can be an integrated charge readout (this charge is equal to $Q_0/2$ for uniform charge deposition in the gap) or a current readout. In the first case, the response is rather slow (~400 ns for a 2 mm gap in LAr). In the latter (Fig. 6.35) the response can be much faster (~40 ns rise time with a suitable CR-RC2 electronics filtering) but the signal to noise ratio is worse given that less "equivalent" charge is sampled, and the bandwidth of the electronics needs to be larger. At high speed (current readout) the limitation comes from the capacitance and inductance of the elementary readout cell, which must be kept appropriately small.

For LHC applications the optimization for high rate requires current readout with fast shaping, together with high granularity to limit pile-up of showers from consecutive events. While the electronics noise decreases when the electronics response becomes slower, the pileup noise generated by low energy particles from consecutive events increases. The shaping time is optimum when the two contributions are equal (see Fig. 6.36). One of the most ambitious realizations is the electromagnetic calorimeter of the ATLAS experiment at the LHC, which uses an 'accordion' geometry [89] to achieve the LHC performance specifications. This geometry provides full azimuthal symmetry without "cracks" between adjacent modules. The geometry, which includes three samplings in depth, is shown in Fig. 6.37. More details about the ATLAS calorimeter are given in Sect. 6.7.4.

The NA 48 collaboration at CERN developed a homogeneous noble liquid ionization calorimeter [90]. It had a cross-section of 2.5 m × 2.5 m, and was optimized for the study of neutral decays of high-energy neutral kaons. Liquid krypton was chosen as compromise between short radiation length (LXe would be preferable) and acceptable cost (the radiation length of argon is too large for fitting

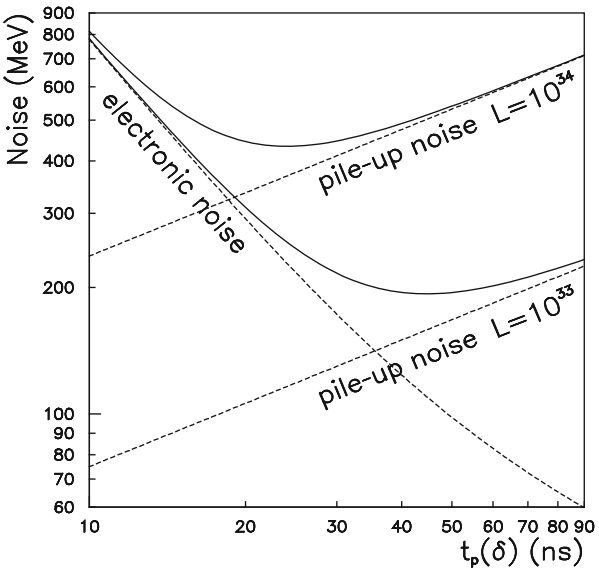


Fig. 6.36 Optimization of shaping time as a function of preamplifier noise and pile-up noise

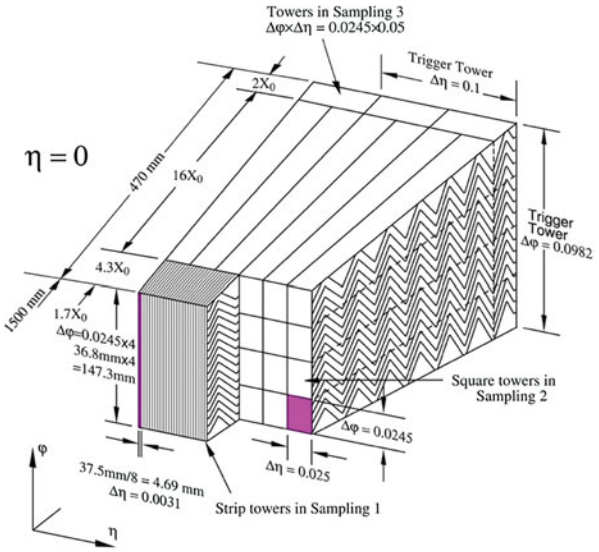


Fig. 6.37 Conceptual view of the 'accordion' geometry

a calorimeter able to contain high energy showers in an acceptable longitudinal space). Readout cells were defined by thin copper-beryllium ribbons stretched in the direction of the beam. The width of the bands (2 cm) and the gap (double gap of 2 × 1 cm) defined readout cells of 2 cm × 2 cm, smaller than the Molière radius

of krypton. In order to smooth the sampling of the shower, the bands were given a zigzag shape in depth by passing the ribbons through staggered glass-epoxy frames. The preamplifiers, connected to each signal band through a blocking capacitor, were located in the liquid for best performances. This calorimeter operated at a high voltage of 3 kV (0.3 kV/mm electric field), in a stable way during several years, with performances characterized by a stochastic term of $3.5\%/\sqrt{E}$, a signal peaking time of 80 ns, a noise per cell of 9 MeV (about 100 cells are needed to reconstruct with high accuracy an electromagnetic shower), a linearity better than 1 part in a thousand between 10 and 90 GeV, and an uniformity of response of 0.5%. Liquid krypton is also being used for the calorimeter (KEDR) of the VEPP2M collider at Novosibirsk [91].

Homogeneous noble liquid calorimeters with very high granularity readout can lead to very interesting imaging and energy measurement properties. One concept, inspired by gaseous tracking chambers (TPCs), was pioneered by the ICARUS collaboration [92, 93]. A more recent example is microBoone at Fermilab [94]. Detectors of this type with long drift distances (1 m or above) find their application in low rate experiments, such as neutrino experiments. The DUNE project, already mentioned, combines the readout of scintillation light and ionization.

A potentially attractive alternative to noble liquids is the use of silicon detectors. However, due to the high cost of silicon diode sensors, the silicon calorimeters operated so far have been restricted to places where the lack of space, and the limited volume, made the use of this technology mandatory. An example is given by SiCal [95], the luminosity calorimeter of the Aleph experiment at LEP. It consisted of a stack of 12 layers of silicon sensors interleaved with tungsten absorber plates, for a total thickness of ~24 X_0 in a longitudinal extension of only 150 mm. High resistivity, n-type (7 kΩcm, 300 μm thickness) Si was used for the 1.3 m^2 readout area, divided into 12,228 channels. The primary purpose of the detector was an absolute measurement of the luminosity using Bhabha scattering. The precision in the reconstructed position of showers (see Sect. 6.4.1) and the precision of the detector acceptance and alignment were essential for the measurement.

For the High-Luminosity LHC phase (HL-LHC) the CMS collaboration is embarking on an extremely ambitious replacement of the electromagnetic part of its end-cap calorimeters. Sampling calorimeters with Si-diode readout are being developed. The total Si readout area will be 600 m2 with a total of 6 million readout and 1 million trigger channels. Remarkably, intensive R&D has demonstrated that the Si detectors will withstand the radiation load [96]. This approach will be taken one step further for detector facilities at future colliders, such as a e^+e^- Linear Collider, with Centre of Mass energy up to several hundreds of GeV. Electromagnetic and hadronic calorimeters with extreme granularity and up to 100 million channels are being considered [97]. For such devices the use of Silicon sensors is one technology of choice. The cost of this option may be an obstacle, to be weighted against the potential performance advantages (see Sect. 6.5). In the forward direction, where the level of electromagnetic radiation from the beams is expected to be high, more radiation resistant sensors, like diamond, are being considered [98].

6.3.4 Gas Detectors

Charge collection in gases, usually followed by some degree of internal amplification, forms the basis of another important category of ionization sampling calorimetry. This method lends itself naturally to highly segmented construction and has profited from the diversified developments of gaseous position detectors (see Chap. 4). The relatively low costs of gaseous detectors favours their use in large area applications such as calorimeters for neutrino physics.

While gaseous ionization calorimetry offers several of the advantages found in ionization calorimetry with dense active materials, the low density of the gaseous readout planes—even compensated by internal charge amplification—limits the performance of such devices [29]. The low density has several disadvantages: Landau fluctuations of the energy deposit in the active gaseous layers can be comparable to the mean deposit and contribute to fluctuations at levels similar to sampling fluctuations; low-energy shower-electrons may multiple-scatter into the readout planes, where they may travel distances large compared to the gap thickness of the active layer, resulting in path-length fluctuations. These effects are relatively unimportant in dense materials, but may reach the level of Landau fluctuations in gaseous readout. Soft particles in the shower will spiral in strong magnetic fields, further increasing these path-length fluctuations. The absolute level of gas amplification depends on external operating conditions (pressure, temperature, gas composition) and is therefore difficult to control precisely. Variations of gas amplification also contribute to worsening the resolution.

An illustration is the electromagnetic calorimeter of the Aleph experiment at LEP [99]. The barrel part of the calorimeter consisted of 12 identical modules surrounding the central tracking system (a Time Projection Chamber), immersed in a solenoidal magnetic field of 1.5 T. The modules had 45 lead/wire-chamber layers for a total of 22 X_0. The cathodes of the readout chambers were segmented into pads of ~30 × 30 mm, providing energy and position information for each shower. The calorimeter was operated with a xenon-CO_2 mixture to increase the density of the active medium, thus reducing pathlength fluctuations. Wires connected to the pads of each layer were brought to module edges, where they were grouped into towers pointing to the vertex. The towers were segmented in three layers in depth of 4, 9 and 9 X_0, respectively. The connections of individual pads to the module edges resulted in a large inductance and therefore limited the rise-time of the readout signals (in the μs range). This was acceptable at LEP given the low event rates. This calorimeter, segmented in 74,000 towers, had an energy resolution of $\sigma(E)/E = 0.18/\sqrt{E} \oplus 1.9\%$, with E expressed in GeV (due to internal amplification, the electronics noise term was negligible).

One of the weak points of this technique is the non-linearity of response. Test beam studies showed that the energy E_{raw} recorded for electromagnetic showers needed to be corrected by:

$$E_{\text{corr}} = E_{\text{raw}} \left(1 + 0.00078\ E_{\text{raw}}\,(\text{GeV})\right),$$

implying a 7.8% correction at 100 GeV. Such non-linearities affect in particular high energy jets in which several showers may be superimposed, thus affecting the result in a way difficult to correct.

While this technique was still adequate at LEP, gas calorimeters were not considered for the LHC. With a very small cell size allowing a binary readout, they may find some application in hadronic calorimetry for the ILC (see for example [100]). An exception at the LHC concerns the very forward region in which, due to the high density of energy deposits, gas ionization chambers (ie without any amplification) are being used for specific purposes, including beam loss monitoring and luminosity measurements [101].

6.3.5 High Rate Effects and Radiation Damage

High particle rates and associated backgrounds impact both on the performance and the useful operating time of calorimeters. Radiation damage needs to be considered for the active readout material and signal processing electronics. Particle rates drive the choice of the calorimeter technology and construction.

Calorimeters with gaseous readout are particularly vulnerable to the high radiation environment due to the ageing effects associated with internal gas amplification, as discussed in Chap. 4.

Such radiation damage is essentially absent in noble liquids making this technology one of the most intrinsically radiation-hard techniques used to date. However, care has to be taken to select adequately radiation resistant components, including electronics, to limit deterioration of the performance (e.g. due to out-gassing). Particularly vulnerable are plastic insulators used in multilayer electrodes or in signal cables. Among the insulators highly resistant to radiation and suitable for calorimeter construction are polyimide (like Kapton) and PEEK. A fundamental limitation of noble liquid calorimeters are space charge effects due to the low drift speed of the positive ions (typically in the range of few cm/s at a nominal electric field around 1 kV/mm). At high incident rates these ions form locally a charged domain which effectively shields the electrons in the gaps from the externally applied field, reducing the drift velocity and thus the signal. These space charge effects are inversely proportional to the square of the detector gaps [102]. For this reason the forward calorimeters [103] of the ATLAS experiment feature gaps down to 250 μm.

Scintillators suffer from the formation of colour centres, which absorb part of the emitted light. The qualification of $PbWO_4$ as a candidate for the CMS crystal calorimeter required a world-wide R&D programme to study the radiation damage effects and to develop methods of crystal growth improving the radiation hardness. Several impurities were identified, which affect transparency in the useful wavelength range (above 350 nm). The best radiation resistance was obtained for crystals grown in Pb/W stoechiometric conditions, with the addition of a small quantity (~100 ppm) of Nb and Y [104]. These crystals showed a light loss of

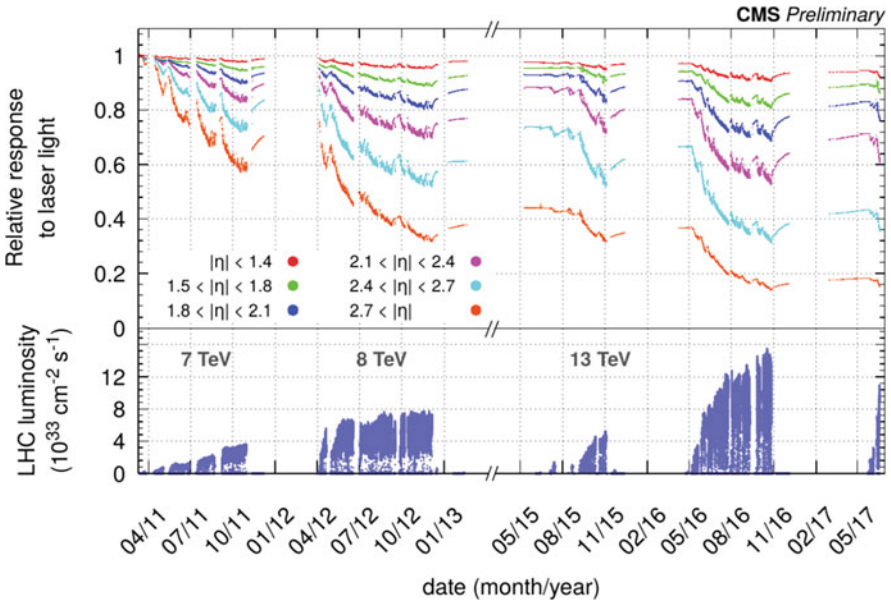


Fig. 6.38 Relative response of the CMS crystal calorimeter to laser light as a function of time, during the initial 5 years of LHC data taking

~3% after an exposure to ~10 Gy in ~10 h, corresponding to the radiation dose accumulated in calorimeters at LHC nominal luminosity during a typical operating period of 20 h. These colour centres show annealing with a recovery time of ~10 h (see also Sect. 3.1.1). After some years of data taking at the LHC, with instantaneous luminosities up to twice the nominal (i.e. 2 10^{34} cm^{-2} s^{-1}) and close to 100 fb^{-1} of accumulated data at 13 TeV in the centre of mass, there is enough experience to judge the crystal behaviour, conveniently followed using laser pulses sent in turn to each crystal. At central rapidities the light loss remains small, due to effective annealing between data taking periods. Some permanent damage accumulates in the more forward region. This is illustrated in Fig. 6.38 [105].

Radiation effects on the light transducers (APD) give an additional contribution to the electronics noise, still rather minor after the integrated luminosity quoted above.

As anticipated, the response of the ATLAS liquid argon calorimeter remains stable during LHC running. Using the position of the Z^0 mass peak reconstructed from electron-positron pairs, a variation of less than 0.05% over the whole 8 TeV data taking period of the "run-I" in 2012 is observed. The peak position is also independent of the mean number μ of collisions per crossing, ie there are no significant rate effects [106] at least up to μ of order 30.

6.3.6 Calibration and Monitoring of Calorimeter Response

Modern calorimetry operates frequently at the 1% accuracy level and requires therefore appropriate calibration methods. An extraordinary effort went into the development and deployment of adequate calibration techniques for the LHC calorimeters. In general, the following tasks have to be performed:

- establishing the absolute scale of response of a calorimeter, averaged over an entire data set
- assessing the uniformity and linearity of response
- monitoring the response as a function of time, locally and globally, in order to correct for time dependent effects, rate effects, aging.

A few examples are discussed below to illustrate each of these tasks.

Energy Scale

(i) *Low energy domain*: one large-scale example is the Superkamiokande experiment, dedicated to low-energy neutrino interactions. After a careful calibration of the gain of each of the phototubes, and an assessment of the water transparency (absorption length greater than 100 m), the absolute energy calibration was made using two radiation sources for cross-checks:

- the beam of an electron Linac operated in-situ above the liquid volume was sent through an evacuated beam pipe into several places of the detector volume recording the corresponding light signals. The Linac was operated at energies between 5 and 20 MeV. The absolute energy scale of the beam was known to better than 1%;
- ^{16}N radioactive nuclei were produced in situ from ^{16}O nuclei of the water volume using a neutron generator. The decay products to $^{16}O^{*}$ (beta emission with an endpoint energy of 4.3 MeV in coincidence with a 6.13 MeV photon) were then recorded during a few lifetimes of ^{16}N (7.13 s). The two methods agreed to better than 0.6% rms.

(ii) *Medium energy domain*: one example is the Babar experiment at SLAC, which used a CsI crystal electromagnetic calorimeter and employed three calibration sources to cover the full energy range:

- at low energy, the 6.13 MeV photons of ^{16}N decays were used (see Superkamiokande above). At this energy, the resolution of the calorimeter was found to be $5 \pm 0.8\%$.
- at high energy (~10 GeV) the Bhabha scattering was used. With a luminosity of $3 \cdot 10^{33}$ cm^{-2} s^{-1} this reaction provided about 200 events per crystal in a 12 h run.
- finally the peak position of known neutral resonances decaying in two photons were used for further checks. Figure 6.39 shows the recorded $\gamma\gamma$ invariant mass spectrum. The π^0 peak was observed at the nominal mass of 135.1 MeV with a width of 6.9 MeV.

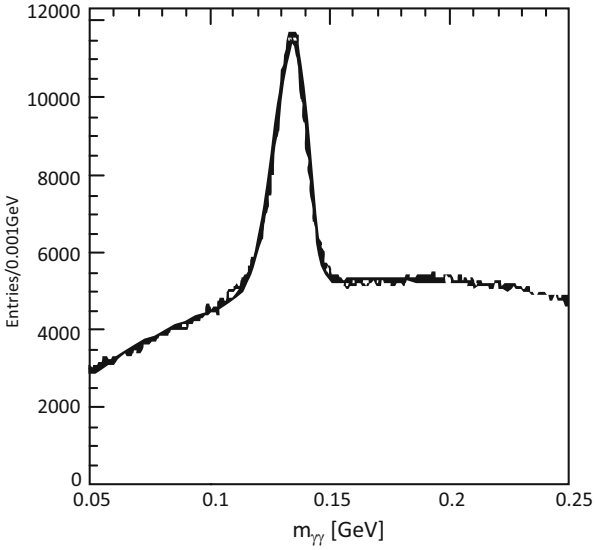


Fig. 6.39 Invariant mass of two photons in $B\bar{B}$ events recorded in Babar. The position of the π^0 peak provides the reference for the energy scale

– Bhabha scattering was used to calibrate the electromagnetic calorimeters of the four LEP experiments.

(iii) *High energy domain*: At the Tevatron the energy scale of the electromagnetic calorimeters was set using the precisely known mass of the Z^0 ($M_Z = 91{,}188 \pm 2$ MeV) decaying into e^+e^- pairs. The LHC experiments rely heavily on this approach given the high rate of Z^0 production: about 10 millions reconstructed Z^0 decays to e^+e^- were used by ATLAS and CMS to establish the energy scale of their electromagnetic calorimeter for the "run-I" at 7 and 8 TeV [106, 107]. The high-accuracy calibration of the electromagnetic calorimeter is essential for precision measurements (at the level of a few tens of MeVs) of the W mass [108] in the $e\nu$ decay mode, and for the measurement of the mass of the recently discovered Higgs boson, using decays in 2 photons, and in 4 leptons [109].

Uniformity and Linearity

With large enough statistics, the Z^0 mass constraint can be used to rescale in situ the response of an LHC calorimeter sector by sector and to improve its uniformity of response. ATLAS uses this method after dividing the calorimeter in about 30 slices in η. The residual non-uniformity is about 0.8% in the barrel region, being somewhat worse (up to about 3% locally) in the end-cap region [106].

If the amount of material in the magnetic spectrometer in front of the calorimeter is low enough, the relation between the energy measured in the calorimeter and the momentum measured in the spectrometer (E/p constraint) can be used to assess both

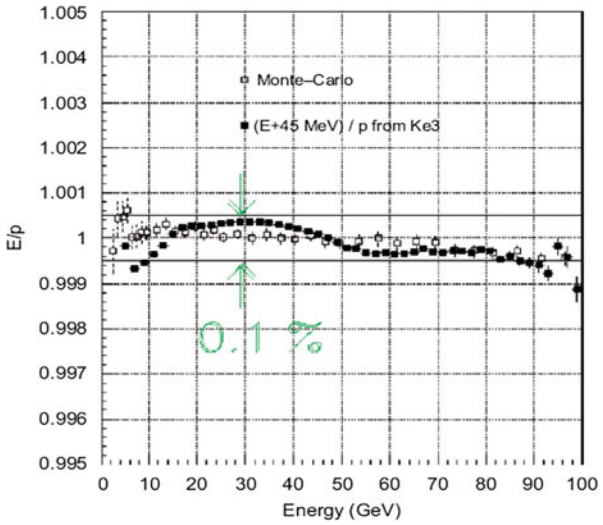


Fig. 6.40 Linearity of the NA48 homogeneous krypton calorimeter. The term added (45 MeV) corresponds to the average energy loss of electrons in the material preceding the sensitive volume

the uniformity and the linearity of response of the calorimeter. A correspondingly high precision mapping of the magnetic field in the spectrometer is of course needed. This technique was used with success in the NA48 experiment with a large sample of Ke_3 decays, demonstrating a linearity better than $\pm 5 \cdot 10^{-4}$ between 10 and 80 GeV, see Fig. 6.40. At the LHC the amount of material in the tracking volume is too large to get the best of this technique. Instead, the large sample of J/ψ decays in electron-positron pairs allows to assess the linearity of the electromagnetic calorimeters between ~5 GeV (high-p_T J/ψ are used in order to have a selective enough trigger) and ~ 50 GeV [107, 110]. An excellent linearity ($\pm$ $1 \cdot 10^{-3}$ between 20 and 180 GeV) was also demonstrated-locally-for ATLAS lead-liquid argon calorimeter modules exposed to a specially equipped beam line at CERN, used as a precision spectrometer (see Sect. 6.7.4).

Monitoring of Short Term Effects

In some cases the calorimeter response is subject to time dependent effects, on a time scale too short to allow for correction with the recorded physics data itself. External monitoring is in this case necessary. An example is the laser monitoring of the CMS crystal calorimeter designed to follow the light absorption and recovery as a function of the instantaneous luminosity, as discussed above, and shown in Fig. 6.38.

In many cases, the detector response depends on operating conditions. As an example, the energy response of the ATLAS liquid argon calorimeter depends on the temperature of the liquid bath with a coefficient of -2% per degree. Precision thermometers (Pt100 resistances) are used to follow the temperature with a precision better than 50 mK. Given the temperature stability observed no short-term correction

was required. In all precision experiments, the gain of the front-end electronics is monitored by injecting precision electrical pulses, allowing subsequent corrections to be made with a precision of 10^{-3} or better.

6.4 Auxiliary Measurements

The analysis of shower properties provides important additional information on position, angular direction and arrival time of the particles which initiated them. Shower shape analysis gives insight on the particle nature. The efforts lavished by the LHC collaborations on electron and muon identification and spectroscopy are eloquent testimony.

6.4.1 Position and Angular Measurements

Conceptually, two methods can be used to obtain spatial information: transverse and longitudinal granularity of the instrument on a scale smaller than the characteristic showers sizes gives position and direction by 'design'. Alternatively, if the readout volume is far larger than the shower dimensions, spatial information may be obtained by 'triangulation' using signals from several sensors distributed over the outer surface of the calorimeter volume.

The latter approach is used for calorimeters with large sensitive volume read out by photomultipliers distributed over their surface (e.g. Superkamiokande). The position is obtained by measuring the difference of light arrival times at the photomultipliers. With a timing resolution between 1 and 3 ns (depending on the pulse height) a position resolution of 70 cm is obtained for 10 MeV showers inside the sensitive volume.

In calorimeters with a more classical tower structure, the position of the incident particle is obtained by calculating the energy-weighted barycentre of energy deposition, using a cluster of cells around the local maximum energy deposition. Because of the finite size of the cells as compared to the Molière radius, the barycentre position is biased towards the centre of the cell with the largest energy deposition. This systematic bias can be corrected by fitting empirical functions. After applying this correction the position accuracy scales as $1/\sqrt{E}$ (decrease of shower fluctuations with increasing energy) convoluted with a constant and a noise term.

In the homogeneous NA48 krypton calorimeter (2×2 cm cells) a position resolution $\sigma_{x,y} = (4.2/\sqrt{E(\text{GeV})} \oplus 0.6)$ mm was measured, while the Babar CsI crystal calorimeter (4x4 cm crystals) gave slightly better results (3.2 mm/$\sqrt{E(\text{GeV})}$. This difference is explained by the smaller Molière radius of CsI (3.8 cm, against 5.5 cm for liquid krypton) and larger signal to noise ratio.

Segmented calorimeters, especially sampling calorimeters with ionization readout, allow lateral and longitudinal segmentation. With two or more samplings in

depth the direction of photon showers can then be estimated. As is shown in Fig. 6.13, the shower is particularly narrow and already well developed after~5 X_0; it is thus advantageous to sample it with high granularity over this depth. In ATLAS, with a cell size of ~5 mm the position of electron and photon showers is determined in the first ~5 X_0 (above~30 GeV) with an accuracy of about 300 μm, a critical asset for physics at the LHC. An important example is the discovery for the Higgs boson using the two-photon final state. The ATLAS electromagnetic calorimeter has three longitudinal samplings for measuring the direction of photons with an accuracy of about 50 mr/$\sqrt{E}$. This angular resolution is such that it makes a negligible contribution to the Higgs mass resolution [111], even if the interaction point cannot be identified among the numerous primary collision vertices at high luminosity. Search for new long-lived neutral particles decaying into photons (like gravitinos) also benefit from a high-resolution angular measurement.

6.4.2 Timing

The electromagnetic cascade develops at the sub-nanosecond timescale, allowing accurate timing measure-ements. This measurement allows identifying the bunch crossing associated to a particular event at colliders. Timing may be used to infer the shower position (see Sect. 6.4.1) or may discriminate between relativistic electromagnetic and slow particles, such as antineutrons.

In a segmented calorimeter the timing resolution is limited by fluctuations of the light path reflecting on edges of the tower, in case of light readout, or by electrical signal reflections at the ends of tower electrodes in case of ionization readout. Electronics noise and shower fluctuations introduce a further limitation, dominant at low and medium energies. While the energy in a tower can be obtained by sampling the signal at its maximum, the optimal time measurement requires additional signal processing. Constant fraction discriminators or digital treatment of multiple samplings of the signal (also beneficial for energy measurements) are frequently used. The shaping time of the electronics is a critical parameter in optimizing the timing accuracy.

As an example, the homogeneous NA48 krypton calorimeter showed a resolution of $\sigma = 0.5$ ns/$\sqrt{E}$, up to ~100 GeV. With the light readout in the "spaghetti" lead-fiber sampling calorimeter of KLOE [82] a spectacular resolution of 0.054 ns/$\sqrt{E} \oplus 0.14$ ns was obtained for photons between 50 and 300 MeV, allowing the shower barycentre along the spaghetti bar structure to be located with a precision of ~3 cm.

With a time resolution better than 100 ps, vertex localisation becomes possible, with an accuracy of a few cm. At the LHC, the rms spread of collision vertices along the beam axis is about 5 cm or ~180 ps. At high luminosity when 50 to 200 collisions per bunch crossing are observed, or envisaged (in the case of HL-LHC), a significantly better resolution is required in order to help in the vertex selection. Upgrade projects at HL-LHC are aiming at 30 ps, which seems the best

possible value with the technology available or under development. One of the most advanced projects is the High Granularity Calorimeter (HGCal) replacement of the crystal system in the endcaps of CMS [96]. In the dense core of the early part of the shower, the signal to noise ratio and the intrinsic shower fluctuations are such that a ~20 ps resolution has been obtained with Si diodes. A similar precision could possibly be reached for non-showering particles (mips) by using "low gain avalanche diodes" as developed and tested by several groups [112, 113].

For hadronic showers, the time development of the energetic component of the cascade is of the order of tenths of nanoseconds, whereas the thermal neutron capture may extend up to 1 μs. Nevertheless, typical time resolutions are found to be at the level of 1–2 ns/$\sqrt{E}$. As an example, with multiple digital sampling a time resolution of $\sigma = 1.5$ ns/$\sqrt{E}$ is measured in the ATLAS Tile Calorimeter [114]. The different time evolution of electromagnetic and hadronic showers offer interesting possibilities for improved shower treatment, a feature likely to be exploited at future facilities (see 6.7.6.2).

6.4.3 Electron and Photon Identification

Apart from certain final states easily identified, like Bhabha scattering at e^+e^- machines, electrons and photons are in general buried inside the copious production of hadrons or jets. This is particularly true at hadron colliders where the electron/hadron ratio ranges from 10^{-3} to 10^{-5}. Since electrons and photons are often signatures of interesting physics, their identification at the trigger and analysis level is crucial. The basic criterion for electromagnetic shower identification is the transverse and longitudinal shower shape, restricting em showers to the electromagnetic compartment, as opposed to hadrons and jets depositing energy in the full calorimeter. This condition is easy to implement, already at the trigger level. Comparing shower shape parameters in the electromagnetic compartment (width, length) to pre-programmed patterns provides the needed additional discrimination. Further discrimination is obtained by treating electrons and photons separately. An electron is signed by a charged track pointing to the shower barycentre, with a momentum p compatible with the calorimetric energy E. The rejection power of this E/p test is however compromised when the electron starts to shower in the tracking device in front of the calorimeter, distorting the momentum measurement and possibly the calorimetric measurement. The remaining background is dominated by π^0s overlapping with a charged pion. A photon is identified through the absence of a track pointing to its barycentre. At this stage the background for photons is often dominated by a π^0 decaying into close-by photons. Very fine granularity in the first ~5 X_0 is one approach to reject these π^0s. As an illustrative figure, simulations made for the ATLAS experiment, give a rejection factor of jets of about 3000 (for a photon acceptance of 80%), when studying the γ + jet final state as a possible background to the $\gamma\gamma$ reaction, with photon energies around 50 GeV [115]. For certain physics reactions an 'isolation criterion'-absence of tracks above a certain

p_T, nor calorimeter energy in a cone around the electromagnetic shower can be applied to sharpen photon or electron identification. This criterion does not apply e.g. for electrons resulting from heavy quark decays inside a heavy quark jet.

The Higgs boson discovery in the di-photon mode was a brilliant demonstration that the necessary jet rejection was achieved by both ATLAS and CMS experiments. At an invariant mass of the Higgs boson of about 125 GeV, the di-photon continuous background consists of about 75% prompt di-photons, 20% photon-jet background and about 5% jet-jet background.

Samples of electrons-positrons with an invariant mass around the Z^0 mass allow a clean measurement of the electron sample purity, as well as of the selection efficiency, using the "tag and probe" method, see Refs. [107, 110] for details.

6.4.4 Muon Identification

The registration of muons in calorimeters contributes to their identification, provides an important means of cross-calibration and in-situ monitoring of calorimeter cells and is used to improve the quality of the muon spectroscopy for instruments located behind the calorimeter.

Identification relies on the reconstruction of a penetrating, charged track behind the hadron calorimeter and possibly on the measurement of an energy deposit in the calorimeter cells along the path of the muon. Typical most probable energy deposits in an electromagnetic calorimeter (e.g. the CMS $PbWO_4$ calorimeter or the ATLAS Accordion) are of order 300 MeV, whereas in the hadronic calorimeters several GeVs are deposited. Such values are in general large compared to electronic noise and to energy deposits from particle background. In the ATLAS hadron calorimeter muons deposit more than ten times the energy from particle background due to average inelastic collisions, even in case of event pile-up at the highest collision rates.

Identification and triggering on muons based on calorimeter information is an essential complement to the main muon trigger using tracking chambers, for physics reactions producing low-p_T muons, e.g. tagging c- or b-jets, or detecting J/ψ or Y; production.

Muons are abundantly produced in pp. collisions (see Fig. 6.41). At low p_T the rate is dominated by 'punch-through' particles, i.e. hadrons, which have not interacted in the calorimeter. At high p_T prompt muons (in particular from W decay) become dominant. [116].

6.5 Jets and Missing Energy

Jet spectroscopy and the related signature of 'Missing Transverse Energy' (MET) have contributed to major discoveries (gluon, W-boson, top quark, ...). At LHC,

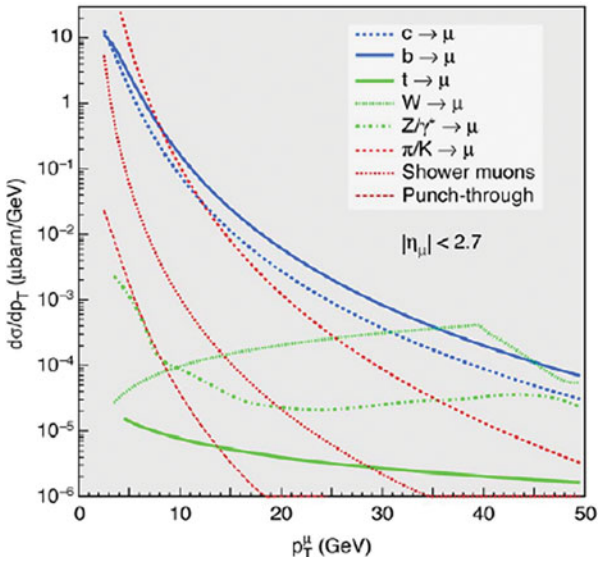


Fig. 6.41 Estimated muon spectra from various sources in the ATLAS Muon Spectrometer

MET is a key signature, e.g. for SUSY and/or dark matter searches. Very high-performance jet spectroscopy is also one of the principal design considerations for future Collider Detectors. The resolution and linearity of the jet energy reconstruction is the principle performance criterion.

The measured jet energy has to be related to the corresponding parton (quark, gluon) energy in a sequence of complex steps. Initial and final state gluon radiation and parton fragmentation affect the observable particle composition and momenta in the jet, limiting the 'intrinsic' parton energy resolution to order $\sigma(E_{\text{parton}})/E \approx 0.5/\sqrt{E_{\text{parton}}(\text{GeV})}$ [117]. Experimental factors—different response as a function of particle species and momentum, nonlinearities, insensitive detector areas, signal noise, magnetic field—require large corrections. Finally, jets are not uniquely defined objects. Different procedures are used to attribute a particle to a given jet. The choice of 'jet algorithms' influences the energy attributed to the jet, as do the additional particles in the 'underlying' event or particles from other collisions, recorded with the jet ('pile-up') [117, 118]. Two classes of jet algorithms have been widely used: The cone-algorithm draws a cone in the η-φ space with radius $R = \sqrt{}[(\Delta\varphi)^2 + (\Delta\eta)^2]$ around a 'seed', an energy deposit above a certain threshold, calculates the total transverse energy $E_{\text{T}} = \sum E_{\text{Tparticles}}$ and the E_{T} position and iterates around the new cone position until a stable result is obtained. This algorithm is sensitive to soft radiation effects; its well-defined jet-boundary however eases corrections due to the underlying event produced in the hadron collision. The k_{T}—algorithm clusters particles according to their relative transverse momenta over the η-φ space, controlled by a size parameter D. This algorithm is theoretically attractive, because in principle infrared and collinear

safe, but results in irregular jet boundaries and complicates the underlying event corrections. Recent work [119] has given rise to an improved version, the anti-k_T algorithm, which is safe against infrared and collinear divergences of QCD, and has regular boundaries. This algorithm is now the "default" of most LHC analyses using jets. Remarkably, despite the complexity and magnitude of the experimental corrections, modern analyses (and Monte Carlos) achieve experimental jet resolutions comparable to (sometimes even better than) the resolution measured for single hadrons: $\sigma(E_{\text{jet}})/E) \approx \alpha/\sqrt{\sum E_{\text{particles}}(\text{GeV})} \oplus \text{c}$, where $\sum E_{\text{particles}}$ represent the energy of the particles associated with the jet and where α is close to the stochastic and c close to the constant term measured for single hadrons [120, 121].

Within a jet, the electromagnetic part—coming mostly from π^0 decays—is better reconstructed than the charged hadrons—mostly $\pi^{\pm}$ and $K^{\pm}$ or long-lived neutral hadrons (K^0_L, n, $\Lambda, \ldots$). While the latter are only detected in the hadronic calorimeter, modern algorithms aim to "replace" charged hadrons reconstructed in the hadronic calorimeter by the associated charged track, whose momentum is better reconstructed than the calorimeter energy. While this individual replacement of particles requires complex algorithms, the procedure has been constantly improved, giving rise to "particle flow" algorithms (see Sect. 6.2.9) which are alternatives to jet reconstruction from calorimeters alone. CMS [122] in general prefers the more performant "particle flow" rather than calorimeter reconstruction. Particle flow is well suited for algorithms analyzing a substructure within jets in view, for example, of distinguishing between jets originating from a high p_T W or Z from quark or gluon jets [113].

The jet energy scale can be experimentally validated studying specific final states in which the jet is balanced by a well measured object, such as γ + jet(s) or Z + jet(s). Another powerful constraint is provided by W's decaying into two jets. A convenient source for identified Ws is the ttbar final state, abundantly produced at the LHC. In the p_T range from 30 GeV to 300 GeV, the linearity of the jet energy scale over the whole angular range is better than 2% in both experiments [123, 124].

The measurement of MET' is the only way to infer the production of neutrinos or weakly interacting SUSY-type particles. It is defined as the negative vector sum of the momentum of all reconstructed objects (leptons, photons, jets) in an event, projected onto the plane transverse to the collision direction. In general, a "soft term" is added corresponding to tracks or energy deposits not associated to the reconstructed objects. At high luminosity, in order to avoid unwanted contributions from pile-up, only tracks are considered, because of their unambiguous association with the corresponding collision vertex. Empirically, a MET resolution of $\sigma(E_{\text{missing}})/E \approx 0.7\alpha/\sqrt{\sum E_{\text{Tparticles}}(\text{GeV})}$ is observed (at low luminosity) for soft collisions with α expressing the stochastic term for single hadron resolution. Calorimetric systems with an acceptance of at least $|\eta|$ ~5 and very good 'hermeticity' are required to achieve this performance.

For events with high p_T jets, at high luminosity and after adequate corrections for the contribution of the underlying event, and of residual pile-up, the resolution is only weakly increasing with the number of collisions during the relevant bunch

crossing, and is comparable to the level of the single hadronic particle resolution [125, 126].

6.6 Triggering with Calorimeters

The ability of calorimeters to provide rapidly (order 100 ns) information on the energy distribution of the collisions products is one of the major assets of this technique. In the very rich trigger 'menu' of the LHC experiments all but muon physics is based on calorimetric triggers at the first trigger level L1. The calorimeter trigger provides a selectivity of ~10^{-3} and reduces the 40 MHz bunch collisions rate accordingly. A 'Sliding Window' technique is used to search for local energy topologies in the $\Delta\eta \times \Delta\varphi$ transverse energy distribution. The optimum window size depends on the particle type (photons, electrons or jets), on their threshold, the depth of the calorimeter included in the sum and possibly luminosity. More complex topologies requiring isolated energy clusters (e.g. triggering on isolated photons or electrons) are also used. The L1 trigger is implemented with dedicated hardware processors. The trigger decision time or "latency" of its response is fixed, and is typically a few μs. The information contained in all detectors is "pipelined" during this time, in such a way that no dead time is generated by the L1 trigger. In subsequent stages, called "high-level-trigger" (HLT) selection criteria and energy thresholds are sharpened with software-based algorithms. The treatment during these phases is asynchronous, and many processors (up to thousands) work in parallel. One of the main challenges with the trigger systems is to allow recording W and Z leptonic decays (i.e. with transverse momenta thresholds below ~30 GeV) for calibration purposes, and for electroweak physics, without saturating the bandwidth of the data acquisition systems. As luminosity increases, refinements are necessary to meet this requirement. MET and B-tagging are part of the overall menu of the HLT, in which of the order of one thousand different conditions are examined in parallel. Triggers on hadronic decay modes of τs, which rely on narrow hadronic jets in the calorimeters are also implemented in HLT. See Ref [127] as example for ATLAS.

In LHCb, which addresses heavy flavour physics in the pseudorapidity range between 2 and 5, the transverse momentum thresholds are much lower, typically 3 GeV for both the electron and the hadron trigger. Such low thresholds are made possible due to the lower luminosity operation of the experiment (typically 0.4 $10^{\mathbf{33}}$ $\mathrm{cm}^{-\mathbf{2}}$ $\mathrm{s}^{-\mathbf{1}}$) and the high data acquisition rate (up to 1 MHz). See Ref [128] for details.

6.7 Examples of Calorimeters and Calorimeter Facilities

The development of calorimetric facilities was and continues to be driven by the main directions of particle physics. Not surprisingly, as particle physics had its origin in cosmic ray studies, rather crude hadronic sampling calorimeters were successfully used to measure the energy spectrum of cosmic rays [52]. Electron scattering experiments provided the impetus for the development of homogeneous [129] and sampling [130] electromagnetic calorimeters. A major step in understanding and perfecting hadronic sampling calorimeters was made for the study of hadron scattering experiments, both with protons and neutrons [131]. The basic properties of these instruments were derived and Monte Carlo studies helped to optimize them [132]. The ISR provided the next motivation for a major development effort [35], providing the basis for the calorimeter facilities at Fermilab, HERA and LHC. In parallel, equally innovative calorimeter developments were and are initiated for astro-particle physics.

The recent series of CP-violation experiments in neutral kaon decay has pushed the requirements for electromagnetic calorimetry (Sect. 6.3.3). The LEP physics program emphasized charged particle spectroscopy and identification, with one notable exception, the L3 electromagnetic BGO crystal calorimeter (Sect. 6.3.1) and U/gas hadron calorimeter. For the Fermilab Collider program general purpose electromagnetic and hadronic calorimeter facilities were developed; facilities with new levels of performance were required for HERA, motivated by the need for precision jet spectroscopy (Sect. 6.7.5).

The LHC physics needs state-of-the-art electromagnetic and hadronic calorimetry, optimized for photons at the 100 GeV scale and for jets at the TeV scale, posing challenging system questions, answered in novel and unconventional ways (Sects. 6.7.3 and 6.7.6.1). The Future Collider physics programmes require further performance improvements, particularly concerning jet spectroscopy, exploiting at the same time the specific operation environment (Sects. 6.7.6.2).

6.7.1 The MEG Noble Liquid Homogeneous Calorimeter with Light Readout

The MEG experiment at PSI [73] is dedicated to the search for lepton flavour violation in muon decays. It aimed at a sensitivity for $\mu \rightarrow e\gamma$ decays of 10^{-13}. This requires an outstanding background rejection (for example of the reaction $\mu \rightarrow e\nu\nu\gamma$), requiring a calorimeter with an excellent energy resolution for ~50 MeV photons and a sub-ns response to cope with the high rate.

The half-cylinder shaped calorimeter is shown in Fig. 6.42. It contains 800 litres of liquid Xenon, and is read out by 846 PMTs, covering approximately 30% of the outside surface of the detector volume.

Fig. 6.42 The MEG homogeneous xenon calorimeter during assembly

The PMTs have K-Cs-Sb photocathodes and silica entrance windows transparent to the peak of light emission (175 nm) of liquid xenon.

The detector was optimized for events with a single photon shower in the volume. An interesting technical feature is the construction of the front wall cryostat using a honeycomb technique for better transparency to photons.

High purity (at the ppb level) of the liquid is necessary to prevent absorption of UV photons by contaminants like oxygen and water. The measured absorption length, more than 3 meters, is much longer than the typical light path from emission to the PMTs. The PMT signals are digitized at 2 GHz with a 12 bit accuracy using custom designed electronics.

The energy scale of the calorimeter is calibrated with photons (17.6 MeV) from the Li(p,γ)Be reaction obtained by sending protons from a Cockroft-Walton source to a Li target close to the calorimeter. In addition, photons from π^0 decays produced by π^- hitting a LiF target are also used, with one photon being measured in the Xe calorimeter, and the other one in an auxiliary NaI crystal matrix.

The relative energy resolution at 50 MeV is $\sigma(E)/E = 1.3\%$. The position resolution is ~6 mm and the timing resolution 64 ps. This excellent performance, made possible with this innovative technique, matched the demanding requirements of the experiment.

An upper limit branching ratio of muons decaying to eγ of 4.2×10^{-13} has been published in 2016 [133], based on the total statistics of 7 10^{14} muons stopped in the target. This is the best limit so far. A plan has been put forward and accepted to pursue the experiment with various improvements, and a higher flux of stopping

muons. The liquid Xenon calorimeter is kept, but the PMTs are replaced by VUV sensitive SiPMs, with a size of 12×12 mm^2, in order to improve the photon energy and position resolution. The prospect is to reach a sensitivity of $5\ 10^{-14}$ [134].

6.7.2 The Xenon 1T Experiment

Xenon1T is the largest and most recent detector of a generation of xenon detectors optimized for the detection of nuclear recoils of very low energy (below 100 keV), as could be produced by the scattering of a WIMP on nuclei (xenon in this case). Observation of such recoils, if they were to be produced, requires high accuracy of the energy measurement and very low background. The detector, operated as a dual phase TPC, is sketched in Fig. 6.43 [135]. The sensitive volume is a vertical cylinder of about 1 m diameter and 1 m height. As described in Sect. 6.2.3 both the primary scintillation signal and the ionization signal are exploited.

The ionization electrons are first drifted to the surface by an electric field generated by a set of Copper rings at a linearly decreasing potential from a grounded grid under the surface to bottom. The field intensity is about 12 kV/m. Right above the surface a somewhat higher field accelerates the primary ionization electrons in such a way that they in turn excite (providing secondary photons) and ionize the

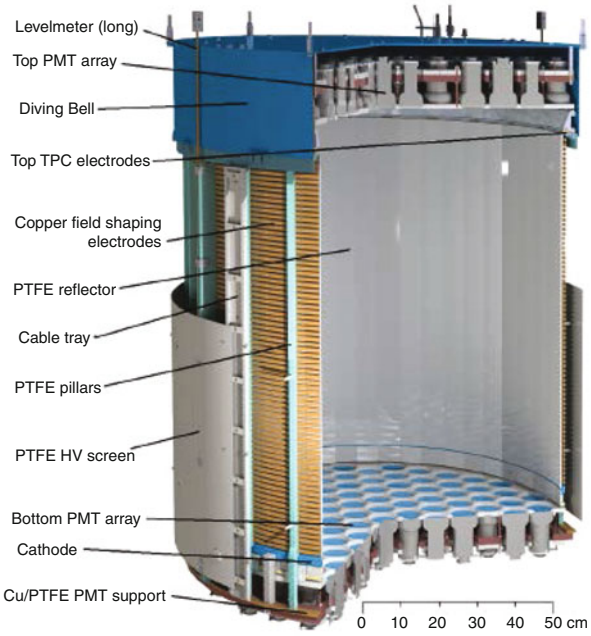


Fig. 6.43 Sketch of the Xenon-1T detector

surrounding gas. Both the primary and secondary photons are detected by a set of 248 VUV photomultipliers, with 78 mm diameter and 35% quantum efficiency at 175 nm, disposed in the liquid at the bottom of the vessel, and in the gas above the multiplication region. The light distribution in the top and in the bottom circles gives the position and lateral extension of the emitted signal. The time between the primary and secondary signals gives the vertical coordinate. All construction materials of the detector were selected for low radioactivity. The experiment is operated in the LNGS laboratory near the Gran-Sasso tunnel, shielded from cosmic background. It is furthermore enclosed in several layers of passive and active shielding. The remaining background is dominated by electron recoils from residual γ emitters, and nuclear recoils from residual neutron background. The former are very much suppressed by a requirement on the ratio of ionization over primary scintillation. The electron lifetime, which depends critically on the extreme liquid purity, and affects the magnitude of the ionization, is measured with photon to electron conversion signals generated in the liquid. A neutron generator is used to calibrate the energy response to recoils. The PMTs and electronics chain are calibrated with blue light pulses sent in fibers ending in the liquid volume. The dark count rate of the PMTs during the first science run was about 10 to 20 Hz. A first science run of about 30 days demonstrated that Xenon1T is the most sensitive device for WIMP masses above 10 GeV presently running. A science run of two years is planned. An enlarged version of the detector, Xenon-nT, with 8 tons fiducial volume is under construction. Its sensitivity should allow to approach the "neutrino floor" given by coherent scattering of solar neutrinos on nuclei.

6.7.3 The CMS Electromagnetic Crystal Calorimeter

The largest crystal calorimeter operated so far is the $PbW0_4$ calorimeter of the CMS experiment at the CERN LHC [110], clearly aimed at the Higgs $\rightarrow \gamma\gamma$ discovery. The calorimeter consists of a cylindrical barrel part (inner radius ~ 1.3 m) and two planar end-caps closing the cylinder at about 3 m from the proton-proton collision point (see Fig. 6.44). Each of the 61,200 barrel crystals is a tapered bar covering a $\Delta\phi \times \delta\eta$ solid angle of 0.018×0.018, and has a depth of 23 cm (24.7 X_0). In the end-caps, the calorimeter is preceded by a lead-Silicon strip preshower. Basic properties of $PbW0_4$ have been given in Sect. 6.3.1.

The calorimeter is located inside the hadronic calorimeter, which in turn is inside the 3.8 T superconducting solenoid. Barrel crystals are readout by APDs, while the end-cap crystals (somewhat bigger) are readout by phototriodes chosen for their better radiation resistance.

The front-end electronics processes signals corresponding to energy deposits of up to ~1.5 TeV (3.0 TeV) in the barrel (end-caps). The equivalent noise per crystal is ~30 MeV. This figure is likely to increase after high luminosity running, due to increased leakage current in the APDs.

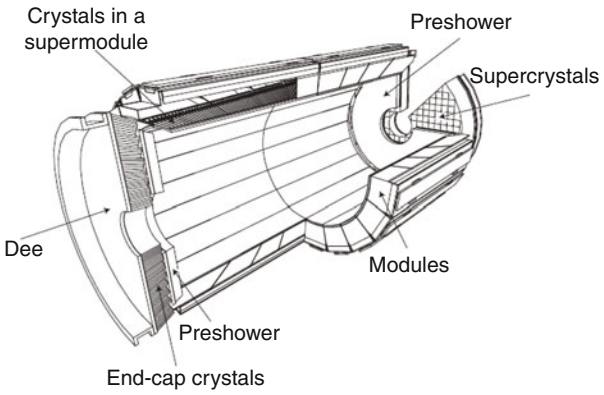


Fig. 6.44 Layout of the CMS electromagnetic calorimeter, showing the arrangement of crystals, with the preshower in front of the end-caps

Despite stringent quality controls during the crystal production, the particle response as observed in beam tests, showed an unavoidable crystal-to-crystal response dispersion of about 7% rms. Two calibration campaigns with beam test and cosmics were undertaken to establish the calibration constants for the initial LHC operation. Using various tools available at the LHC, like azimuthal uniformity of response, π^0, J/Ψ and Z^0 invariant mass constraints, all crystals were quickly intercalibrated to a precision around 1%. The laser pulse system monitors the short term response variations due to radiation effects.

The CMS crystal calorimeter successfully achieved its essential role for the experiment, both for triggering, as the source of identification and precise measurement of electrons and photons, and as input to particle flow. Among the most important results, based in particular on the calorimeter data, is the already mentioned discovery of the Higgs boson in 2012, revealed in the inclusive di-photon spectrum shown in Fig. 6.45.

6.7.4 The ATLAS Liquid Argon Electromagnetic Calorimeter

While ATLAS and CMS have almost identical physics programs, with the search for the Higgs boson as one of the main objectives, the two experiments have opted for a series of different detection techniques. The ATLAS electromagnetic calorimeter [103] uses a lead/liquid argon sampling technique, with an 'accordion' geometry, and is located outside of the inner solenoid. The liquid argon technique was chosen for its immunity to radiation, its intrinsic stability and linearity of response, and its relative ease of longitudinal and transverse segmentation. Its more modest intrinsic resolution is a limiting factor at medium and low energies.

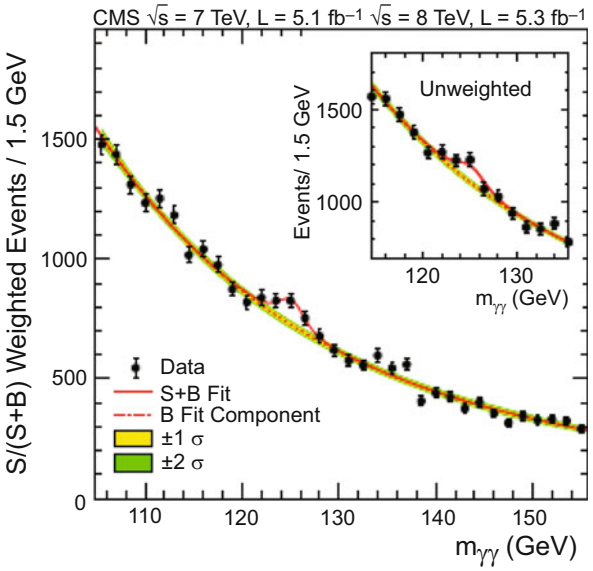


Fig. 6.45 Inclusive di-photon mass spectrum in CMS, from the Higgs discovery paper

The calorimeter features three segments in depth, the first one having an extremely fine segmentation in pseudorapidity (0.003) to allow separation between prompt photons and photons from π^0 decays up to $p_T \sim 70$ GeV/c, the interesting range for the Higgs boson search in the $\gamma\gamma$ decay mode.

The calorimeter is preceded by a presampler, located in the same cryostat, to correct for the loss of energy of electrons and converted photons in the inner detector material, in the solenoid and cryostat front walls (see Table 6.5). The barrel part, consisting of two cylinders, and the two end-cap wheels provide uniform azimuthal coverage despite being built of 16 (8) modules per cylinder (wheel) (Fig. 6.46).

The front-end electronics was optimized (Fig. 6.36) for best performance at the nominal LHC luminosity of 10^{34} cm^{-2} s^{-1}. The dynamic range is covered with three channels with gains in the ratio 1/9/81, digitized with 12 bit resolution. In this way quantization noise remains small compared to the noise level after the preamplifier (10 to 50 MeV depending on the sampling) up to the highest expected energy deposition per cell (~3 TeV). Trigger towers of size $\Delta\eta \times \Delta\varphi = 0.1 \times 0.1$ are built by analogue summing of signals at the front-end level, followed by digitization at 40 MHz with 10 bits ADCs (sensitivity of 1 GeV per count).

The uniformity of response within one module and the reproducibility from module to module were checked in a test beam. The overall dispersion of energy measurements in 3 barrel modules and 3 end-cap modules was respectively 0.43% and 0.62% [136]. The local energy resolution was found to be about 1% (rms) at 120 GeV [94], and is well described by $\sigma(E)/E = 10\%/\sqrt{E} \oplus 0.25/E \oplus 0.003$. The energy scale (Sect. 6.3.6) and the long range uniformity have been assessed in situ using the Z mass constraint. An overall "constant term" of about 0.8% in

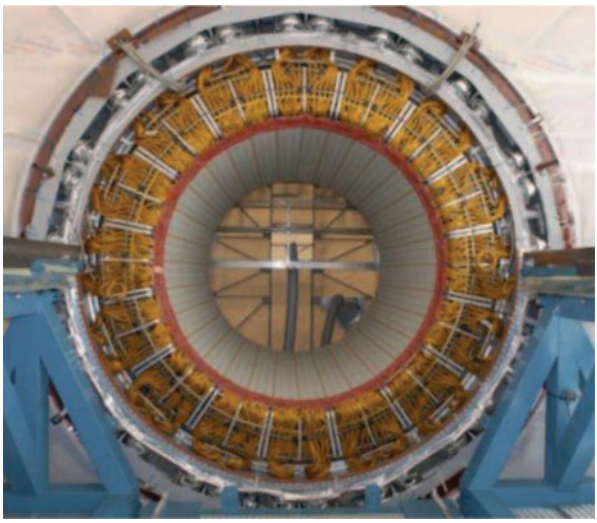


Fig. 6.46 Photograph taken during the assembly of the ATLAS electromagnetic barrel calorimeter. The pre-sampler sectors (in gray) are visible in front of the 16 calorimeter modules

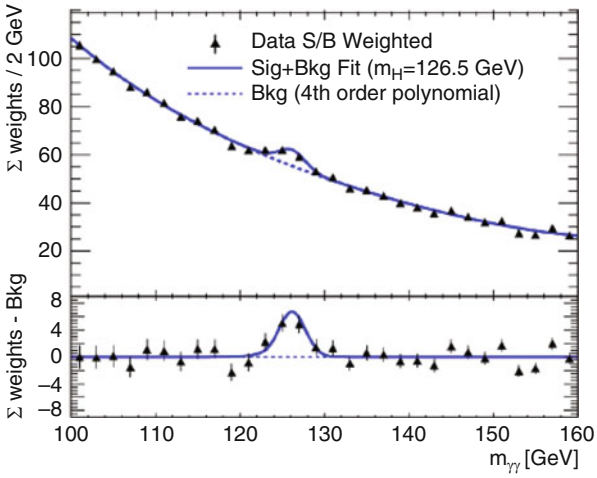


Fig. 6.47 Inclusive diphoton mass spectrum in ATLAS, from the Higgs discovery paper

the barrel and up to 3% in some pseudorapidity ranges of the end-caps covers the unavoidable dispersion in materials and in calibration, and the effects of material in front of the calorimeter not fully described in the simulation. Like for CMS, the electromagnetic calorimeter of ATLAS fulfilled successfully its task. Among the most important results, based in particular on the calorimeter data, is the already mentioned discovery of the Higgs boson in 2012. The corresponding inclusive diphoton spectrum is shown in Fig. 6.47. Also worth mentioning is the contribution of

the electron channel to the recent measurement of the W-mass, 80,370 ± 19 MeV in the muon and electron channels together [137].

6.7.5 *The ZEUS Calorimeter at HERA*

Research at the electron-proton collider HERA required precision jet spectroscopy at the 100 GeV level to study the underlying dynamics of e-quark collisions. Energy and position resolution for jets were at a premium.

The H1 Collaboration developed a calorimeter based on the LAr-Pb and LAr-Fe sampling technology. A certain level of 'off-line' compensation was achieved because hadron showers were measured longitudinally up to ten times and longitudinal shower-weighting could be applied [139].

The ZEUS Collaboration at HERA developed an intrinsically compensated calorimeter using the U-scintillator sampling technique [43, 138], modeled after the Axial Field Spectrometer facility [140]. The calorimeter is constructed in a modular form (Fig. 6.48), with units which are approximately 5 m long, 20 cm wide and more than 2 m deep. The ratio of the thickness of the ^{238}U plates (3.3 mm) to the scintillator plates (2.6 mm) was tuned to achieve $e/\pi = 1$, confirmed by measurements to be $e/\pi = 1.00 \pm 0.03$. The measured hadronic energy resolution, $\sigma\ (E)/E$ (hadrons) $= 0.35/\sqrt{E}$(GeV), is consistent with a sampling resolution of σ/E (sampling, hadrons) $\approx 0.29/\sqrt{E}$(GeV) and an intrinsic resolution of σ/E (intrinsic,

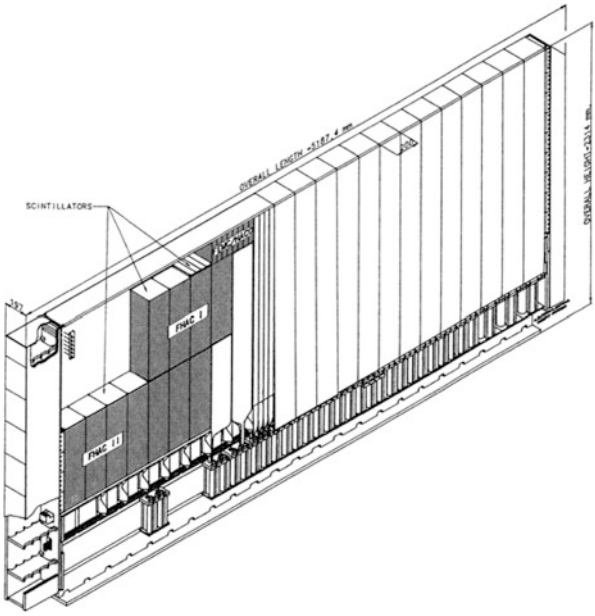


Fig. 6.48 View of a module of the ZEUS U-scintillator calorimeter. Wavelength-shifter readout is used to read cells of $5_{*}20$ cm^2 cross-section in the electro-magnetic compartment and of $20_{*}20$ cm^2 in the two subsequent hadronic compartments [138]

hadrons) $\approx 0.20/\sqrt{E}$(GeV). This sampling frequency is rather coarse for electrons resulting in an electron energy resolution σ/E (electrons) $= 0.18/\sqrt{E}$(GeV).

H1 and Zeus provided a detailed measurement of electron-nucleon scattering from which a new generation of parton distribution functions (PDFs) was derived. These functions have been used, and are still being used extensively for LHC physics analysis.

6.7.6 Facilities at the LHC and a Future Collider

The research programmes at the LHC and at a possible future Colliders impose a new level of performance requirements.

6.7.6.1 Facilities at LHC

The two general-purpose p-p experiments, ATLAS and CMS, have developed rather different approaches for the same physics research, promoted by different groups of physicists with their personal experience, background and taste, constrained by realities of funding. In both cases the extraordinary requirements on electromagnetic calorimetry imposed 'hybrid' solutions to allow independent optimization of electromagnetic and hadronic calorimetry. This 'independence' led ATLAS to choose two novel, unconventional detector geometries. The 'Accordion' calorimeter (see Sect. 6.7.4) is followed by a hadronic instrument with scintillator tile/WLS fibre readout. One of the 64 slices forming a complete and crack-less cylinder is shown in Fig. 6.49. The unconventional geometry of absorber plates and scintillating tiles oriented along the direction of the incident particle permits an economic construction and homogeneous sensitivity [141]. This geometry works because the preceding ~1.5 λ Accordion calorimeter provides enough hadronic shower development to permit good sampling in the Tile-geometry. This arrangement also greatly facilitates longitudinal and transverse segmentation hence permitting effective longitudinal weighting of the shower energies. Weighting leads to a resolution of the combined calorimetry system (accordion and Tile calorimeter) of $\sigma/E \approx (0.52/\sqrt{E} \oplus 1.6/E) \oplus 0.03$ and a good linearity of response [120]. A jet energy resolution of $\sigma(\text{jet})/E \approx 0.6/\sqrt{E}$(GeV) is estimated, adequate for LHC. The ATLAS Tile and Extended Tile calorimeter covers $|\eta| < 1.4$. For the forward ('endcap') regions ($1.4 < \eta < 3.2$) ATLAS had to adopt different solutions to cope with the even more ferocious radiation levels. An Accordion-type electromagnetic calorimeter precedes a Cu/Liquid Argon hadron calorimeter. In the very forward region ($3 < \eta < 5$) yet another novel geometry had to be invented: cylindrical readout elements with narrow LAr-gaps (0.25 to 0.35 mm) as sensitive medium are embedded in a tungsten absorber, sampling geometrically very tight showers at adequate readout speeds [120]. Figure 6.50 shows a cut-view through the ATLAS calorimeter facility.

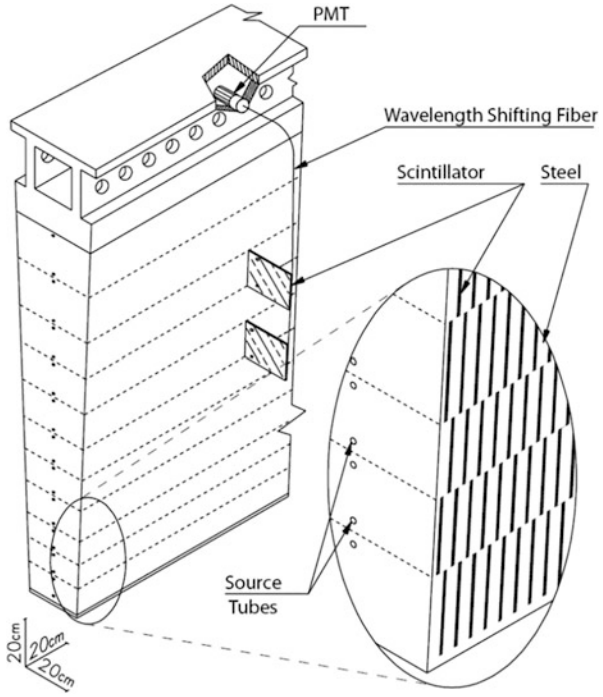


Fig. 6.49 View of one module of the ATLAS hadronic barrel calorimeter. Sixty-four such modules complete the cylindrical detector. Each of the longitudinally oriented scintillating tiles is instrumented with two wavelength-shifting fibers [141]

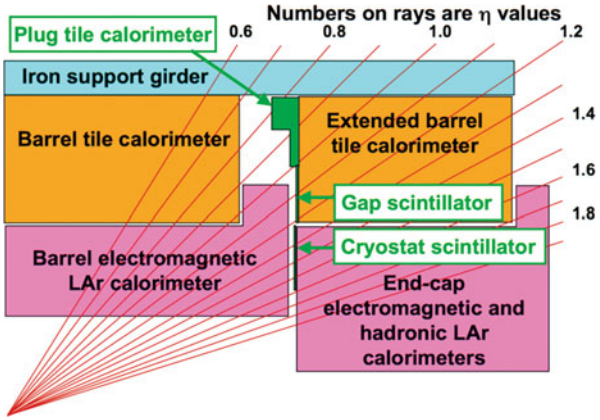


Fig. 6.50 Longitudinal quarter view of the Atlas calori-meter facility. The outer radius is at 4.2 m; it extends along the beam direction to ± 7 m. Auxiliary instru-mentation in the gap between the calorimeters allows energy correction for the non-instrumented zones [120]

Table 6.4 Parameters of the ATLAS and CMS electromagnetic calorimeter facilities

	ATLAS		CMS	
Technology	Lead/LAr accordion		$PbWO_4$ scintillating crystals	
	Barrel	**End-caps**	**Barrel**	**End-caps**
ηcoverage	0–1.475	1.4–3.2	0–1.48	1.48–3
Channels	110,208	63,744	61,200	14,648
Granularity ($\Delta\eta \times \Delta\varphi$)				
Pre-sampler	0.025 × 0.1	0.025 × 0.1	–	–
Strips/Si-preshower	0.003 × 0.1	0.003–0.006 × 0.1	–	32 × 32 Si-strips per 4 crystals
Main sampling	0.025 × 0.025	0.025 × 0.025	0.017 × 0.017	0.018 × 0.003 to 0.088 × 0.015
Back	0.05 × 0.025	0.05 × 0.025		
Depth				
Pre-sampler	10 mm	2 × 2 mm	–	–
Strips/Si-preshower	~4.3 X_0	~4.0 X_0	–	~3 X_0
Main sampling	~16 X_0	~20 X_0	~26 X_0	~25 X_0
Back	~2 X_0	~2 X_0	–	–
Energy resolution				
Stochastic term	10%	10–12%	3%	5.50%
Local constant term	0.20%	0.35%	0.50%	0.50%
Noise per cluster [MeV]	250	250	200	550

CMS calorimetry consists of the novel $PbWO_4$ electromagnetic calorimeter (Sects. 6.3.1 and 6.7.3) followed by a brass (70% Cu, 30% Zn) (50 mm thick) plate/scintillator tile calorimeter. The tiles are optically grouped into towers (0.087 × 0.087 in η-φ space in the barrel calorimeter) and read by hybrid photodetectors, all located in front of the 3.8 T superconducting solenoid. This favourable geometry, however, only allows for a total of ~7 λ, requiring a 'tail catcher' formed by scintillator tiles outside the coil in the first muon absorber layer [142]. Tables 6.4 and 6.5 summarizes the principal design parameters of the ATLAS and CMS Calorimeter Facilities.

6.7.6.2 Developments for Future Collider Calorimetry

The proposal for a future Linear e^+ e^- collider (LC) has triggered a worldwide R&D programme for the appropriate detector technologies [143]. One direction of present R&D addresses calorimetry optimized for its physics programme, emphasizing precision electromagnetic calorimetry and very high granularity for 'Particle Flow Analysis' (see Sect. 6.2.9).

Table 6.5 Parameters of the ATLAS and CMS hadronic calorimeter facilities

	ATLAS	CMS
Technolgy [ηhalf-coverage]		
Barrel/Ext. barrel	14 mm Fe/3 mm scint. [0–1.4]	50 mm brass/4 mm scint. [0–1.4]
End-caps	25 mm (front) − 50 mm (back) Cu/8.5 mm LAr [1.4–3.2]	80 mm brass/4 mm scint. [1.4–3.0]
Forward	Cu (front) − W (back)/0.25–0.50 LAr [3.2–4.9]	4.4 mm steel/0.6 mm quartz [3.0–5.0]
Channels		
Barrel/Ext. barrel	9852	2592
End-caps	5632	2592
Forward	3524	1728
Granularity ($\Delta\eta \times \Delta\varphi$)		
Barrel/Ext. barrel	0.1 × 0.1 to 0.2 × 0.1	0.087 × 0.087
End-caps	0.1 × 0.1 to 0.2 × 0.2	0.087 × 0.087 to 0.35 × 0.028
Forward	0.2 × 0.2	0.175 × 0.175
Longitudinal samplings		
Barrel/Ext. barrel	Three	One
End-caps	Four	Two
Forward	Three	Two
Absorption lengths		
Barrel/Ext. barrel	9.7–13.0	5.8–10.3
		10–14 (with coil/HO)
End-caps	9.7–12.5	9.0–10.0
Forward	9.5–10.5	9.8

One promising direction is being pursued by the DREAM Collaboration [144]. DREAM ('Dual REAdout Method') is a concept aiming at event-by-event separate detection of the electromagnetic component through Cherenkov light and the hadronic showers through scintillation light. Timing information might provide an additional handle to disentangle the various processes (e.g. delayed nuclear photon emission). The combined information could in principle allow complete reconstruction of the shower- and jet composition. The LC jet benchmark resolution of $\sigma/E \approx 0.30/\sqrt{E}$ might not remain a dream [84].

The CALICE (Calorimeters for the Linear Collider Experiment) Collaboration aims at the same performance: it makes the concept of Particle Flow Analysis an integral part of the design concept of the experimental facility aiming to separately measure the momenta of the charged component, photons in the electromagnetic and neutral hadrons (n, K^0) in the hadron calorimeter. The calorimeter is placed at a relatively large radius allowing the jets to open and charged and neutral particles to separate in the strong B-field. This strategy requires exceedingly high granularity (more than 10^8 channels) to measure the individual shower profiles [65].

Besides studies for a possible LC a vigorous programme has been intiated to undertand the physics potential and consequences for experimentation at a possible

"Future Circular Collider" (FCC). A center-of-mass energy for proton-proton collisions in the 100 TeV regime is envisaged, implying a collider circumference of about 100 km. The physics research determines the peak luminosity of about 3.10^{35} cm^{-2} s^{-1}. These key parameters shape the detector design and performance specifications, which are intensively studied [145]. The electromagnetic and hadronic calorimetry emphasizes very high granularity to cope with particle multiplicity and event pile-up, tight control of systematic effects (small constant term), very good linearity and—unsurprisingly—taming the ferocious radiation environment. The calorimeters are of the sampling type, because the stochastic term in the calorimeter performance is less an issue, given that the typical energy scales are in the TeV regime. Simulations show that rather conventional, LHC type calorimeter instrumentation will deliver the desired performance, without excluding novel developments with more "aggressive" technologies. LAr is the technology of choice, except for a possible scintillator option for the central hadron calorimetry. As an indication, the EM calorimeter could be a Pb/ LAr device, with cells sizes between 6_*6 mm^2 to 20_*20 mm^2 and an eightfold longitudinal subdivision. A possible geometry is shown in Fig. 6.51. Hadron calorimtry could be a scintillator/Pb/steel detector (in the central region), which would give e/h $\approx$ 1.1, resulting in the required good linearity and decent jet resolution, see Fig. 6.52.

While these concepts seem plausible, a closer look shows that the technical challenges are formidable . . . fortunately, the LHC experience provided training, motivation and encouragement.

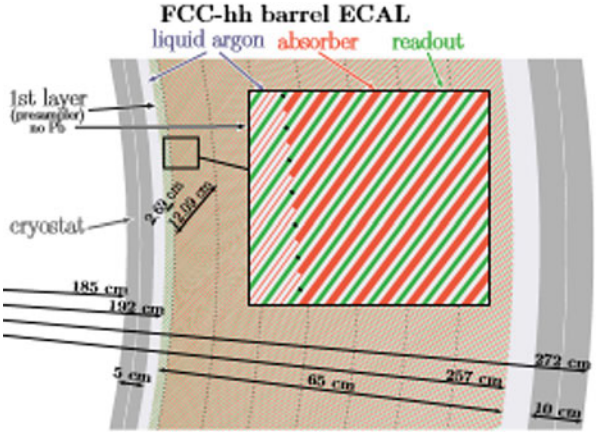


Fig. 6.51 Conceptual structure of an em calorimeter, showing the slanted absorber plates, LAr gaps and readout boards

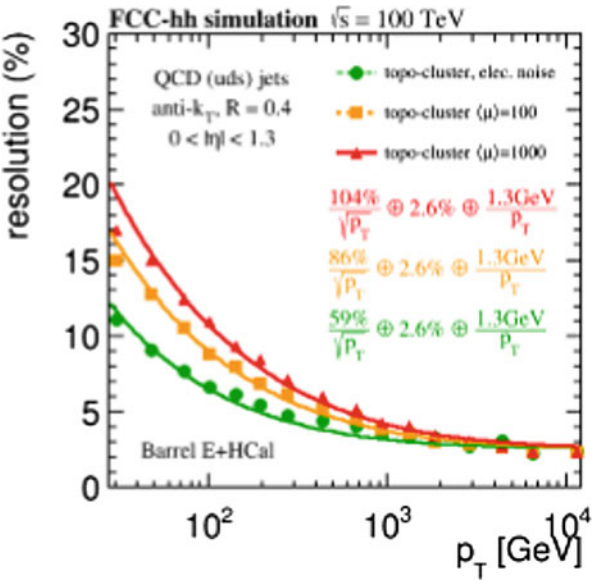


Fig. 6.52 Jet resolution for different hadron calorimeter configurations

6.8 Conclusions

During the past 40 years calorimetry has matured into a precision measurement technique, indispensable to modern particle physics experiments. The Higgs boson, cornerstone of our present understanding of matter, owes its discovery to calorimetry.

Understanding and modelling the physics processes at work in calorimetry at the 1% level has been achieved. Based on this understanding and helped by modern signal processing techniques, developments aim at characterizing the individual showers, at optimizing further particle identification and at reaching the intrinsic performance level for jet spectroscopy, needed for the next generation of precision and discovery experiments.

Acknowledgments We thank many colleagues for discussions and T. Carli for a careful reading of the manuscript.

References

1. M. Tanabashi et al., Particle Data Group, Phys. Rev. D98, 03001 (2018).
2. C. Grupen, Particle detectors, Cambridge University Press (1996).
3. W. Heitler, The Quantum Theory of Radiations (Oxford Clarendon Press) 1954.

4. H.A. Bethe, W. Heitler, Proc. R. Soc. A 146 (1934) 83.
5. Y.S. Tsai, Rev. Mod. Phys. 46 (1974) 815.
6. High Energy Particles, B. Rossi, Prentice-Hall (1952) revised (1961).
7. J.W. Motz, H.A. Olsen, H.W. Koch, Rev. Mod. Phys. 41 (1969) 581.
8. O. Klein, Y. Nishina. Z. Phys. 52 (1929) 853.
9. L.D. Landau, I.J. Pomeranchuk, Dokl. Akad. Nauk SSSR 92 (1953) 535, 735.
10. S. Klein, Rev. Mod. Phys. 71(1999) 1501.
11. R. Moore et al., Nucl. Instrum. Meth. B 119 (1996) 149.
12. D. Schildknecht, *Vector Meson Dominance*, arXiv:hep-ph/0511090.
13. T.H. Bauer, R.D. Spital, D.R. Yennie, Rev. Mod. Phys. 50, no. 2 (1978) 261.
14. H. Burkardt et al., Phys. Lett. B 206 (1988) 163.
15. B. Rossi, K. Greisen, Rev. Mod. Phys. 13 (1941) 240;
16. H.S. Snyder, Phys. Rev. 76 (1949) 1563.
17. W.R. Nelson, H. Hirayama, D.W.O. Rogers, *The EGS4 Code System*, SLAC-265 (1985).
18. S. Agostinelli et al., Nucl. Instrum. Meth. A 506 (2002) 250.
19. E. Longo, I. Sestili, Nucl. Instrum. Meth. 128 (1975) 283.
20. J.C. Philippot, IEEE Trans. Nucl. Sci. NS-17, no. 3 (1970).
21. A. Owens, Nucl. Instrum. Meth. A 238 (1985) 473.
22. U. Fano, Phys. Rev. 72 (1947) 1.
23. F. Gao et al., Nucl. Instrum. Meth. B 255 (2007) 286.
24. H.J. Crawford et al., Nucl. Instrum. Meth. A 256 (1987) 47.
25. E. Aprile et al. Phys.Rev. B76 (2007) 014115
26. J.B. Birks, Phys Rev. 86 (1952) 569.
27. A. Menchaca-Rocha et al., Nucl. Instrum. Meth. A 438 (1999) 322.
28. S. Amoruso et al., Nucl. Instrum. Meth. A 523 (2004) 275 and references quoted therein.
29. U. Amaldi, Physica Scripta 23 (1981) 409.
30. C.W. Fabjan, in: Experimental Techniques in High-Energy Physics, (ed. T. Ferbel), Addison Wesley, Menlo Park (1987).
31. R. Wigmans, Nucl. Instrum. Meth. A 259 (1987) 389.
32. A. Ferrari, private communication (2001).
33. T.A. Gabriel et al., Nucl. Instrum. Meth. A 338 (1994) 336.
34. D.E. Groom, in: Proc. seventh Intern. Conf. on Calorimetry in High Energy Physics, Tucson, AZ, (ed. E. Chen), World Scientific, Singapore (1998).
35. C.W. Fabjan, W.J. Willis, in: Proceedings of the Calorimeter Workshop, Batavia, (ed. M. Atac), FNAL, Batavia, Ill. (1975) 1.
36. R.K. Bock et al., Nucl. Instrum. Meth. 186 (1981) 533.
37. M.G. Catanesi et al., Nucl. Instrum. Meth. A 260 (1987) 43.
38. H. Abramowicz et al., Nucl. Instrum. Meth. 180 (1981) 429.
39. R. Wigmans, Calorimetry: Energy Measurement in Particle Physics, second ed.,Clarendon Press, Oxford (2017).
40. T. Carli et al., arXiv: 1604.01415v2 (2016).
41. C.W. Fabjan, R. Wigmans, Rep. Prog. Phys. 52 (1989) 1519.
42. R. Wigmans, Nucl. Instrum. Meth. A 265 (1988) 273.
43. G. Drews et al., Nucl. Instrum. Meth. A 290 (1990) 335.
44. L. Landau, J. Phys. (Moscow) 8 (1944) 201.
45. P.V. Vavilov, Zh. Eksp. Teor. Fiz. 32 (1957) 920.
46. W. Lohmann et al., CERN-85-03 (1985).
47. Z. Ajaltouni et al., Nucl. Instrum. Meth. A 388 (1997) 64.
48. A. Van Ginneken, Nucl. Instrum. Meth. A 362 (1995) 213.
49. D.E. Groom, N.V. Mokhov, S.I. Striganov, Atomic and Nuclear Data Tables 78, p. 183.
50. W. Lustermann, CALOR2008 (2001).
51. E.B. Hughes, IEEE Trans. Nucl. Sci. 19 (1972) 126.
52. V.S. Murzin, Prog. Elem. Part. Cosmic Ray Phys. 9 (1967) 247.
53. J. Ranft, Nucl. Instrum. Meth. 81 (1970) 29.

54. T.A. Gabriel, J.D. Amburger, Nucl. Instrum. Meth. 116 (1974) 33.
55. Hadronic Shower Simulation Workshop, Batavia, Ill., AIP Conf. Proc. 896 (2006).
56. G. Battistoni, in: Hadronic Shower Simulation Workshop, Batavia, Ill., AIP Conf. Proc. 896 (2006) 31.
57. J. Ranft, in: Hadronic Shower Simulation Workshop, Batavia, Ill., AIP Conf. Proc. 896 (2006) 102.
58. H.W. Bertini, P. Guthrie, Nucl. Phys. A 169 (1971).
59. H.C. Fesefeldt, Univ. Aachen, Report PITHA 85–02 (1985) .
60. C. Zeitnitz, T.A. Gabriel, The GEANT-Calor Interface User's Guide, ORNL, Oak Ridge (1996).
61. M. Aaboud et al., (ATLAS Collaboration),Phys. Rev. D 96 (2017) 077002.
62. M. Aaboud et al., (ATLAS collaboration), Eur. Phys. J. C 77 (2017) 466.
63. F. Beaudette (CMS Collaboration), (2014) arXiv:1401.8155 [hep-ex].
64. M. A. Thomson, in: Hadronic Shower Simulation Workshop, Batavia, Ill., AIP Conf. Proc. 896 (2006), pp. 215
65. C. Adloff et al., arXiv:0707.1245 and http://polywww.in2p3.fr/flc/calice.html
66. A. Fasso et al., Nucl. Instrum. Meth. A 332(1983)459.
67. E. Gschwendtner et al., Nucl. Instrum. Meth. A 482 (2002) 573.
68. F. Sommerer et al., Phys. Med. Bio. 51 (2006) 4385.
69. K. Deiters et al., Nucl. Instrum. Meth. A 453 (2000) 223.
70. B. Dolgoshein et al., Nucl. Instrum. Meth. A 563 (2006) 590.
71. E.Garutti et al., JINST 6 (2011) C10003.
72. N. McKinsey et al., Nucl. Instrum. Meth. A 516 (2004) 475.
73. R. Sawada, Nucl. Instrum. Meth. A 581 (2007) 522
74. B.N.V. Edwards et al. arXiv 1710. 11,032.
75. R.Acciari et al. arXiv:1601.02984.
76. D.S. Ayres et al. (NOvA collaboration), (2005) arXiv: hep-ex/0503053(2005).
77. R.L. Garwin, Rev. Sci. Instrum. 31 (1960) 1010.
78. B. Barish et al., IEEE trans. Nucl. Sci. NS-25 (1978) 532.
79. J. Badier et al., Nucl. Instrum. Meth. A 348 (1994) 74.
80. R. Wigmans, Nucl. Instrum. Meth. A 315 (1992) 299;
81. R.-D. Appuhn et al., Nucl. Instrum. Meth. A 386(1997) 397.
82. F. Ambrosino et al., Nucl. Instrum. Meth. A 598 (2009) 239.
83. OPAL Collaboration, Nucl. Instrum. Meth. A 305 (1991) 275.
84. S. Lee et al., Nucl. Instrum. Meth. A866 (2017) 76.
85. S. Chatrchyan et al., JINST 3 (2008) S08004.
86. S. Fukuda et al., Nucl. Instrum. Meth. A 501 (2003) 418.
87. W.J. Willis, V. Radeka, Nucl. Instrum. Meth. 120 (1974) 221.
88. V. Brisson et al., Nucl. Instrum. Meth. 215 (1983) 79.
89. D. Fournier, ECFA report 90–133(1990) 356, CERN 90/10.
90. V. Fanti et al., Nucl. Instrum. Meth. A 574 (2007) 433.
91. V.M.Aulchenko et al., Nucl. Instrum. Meth. A 394 (1997) 35.
92. S. Amerio et al., Nucl. Instrum. Meth. A 527 (2004) 329.
93. A. Ankowski et al., arXiv:0812.2373 (2008).
94. R.Acciarri et al., JINST 12 (2017) P02017.
95. D. Bederede et al., Nucl. Instrum. Meth. 365 (1995) 117.
96. CMS collaboration, Technical Design Report CMS-TDR-17-007 (2017)
97. J.C. Brient, Proc. XIII Conference on Calorimetery, CALOR2008, Pavia, 2008.
98. R.S. Wallny et al., Nucl. Instrum. Meth. A 582 (2007) 824.
99. D. Buskulic et al., Nucl. Instrum. Meth. A 360 (1995) 481.
100. I. Laktineh, Proc. TIPP09 Conf. (2009).
101. J.F.Beche et al., Proc. IEEE NSSMIC 2000.949101 (2000).
102. J. Rutherfoord, Nucl. Instrum. Meth. A 482 (2001) 156.
103. ATLAS Collaboration, Technical Design Report, CERN-LHCC/96–41.

104. A. Annenkov et al., Nucl. Instrum. Meth. A 426 (1999) 486.
105. Performance of the CMS precision electromagnetic calorimeter at the LHC run-II Z.Zhang, Talk at the CHEF Calorimeter Conference (Lyon-2017).
106. ATLAS Collaboration, Eur. Phys. J. C74 (2014) 10, 3071 and arXiv 1407.5063.
107. CMS Collaboration, JINST 8 (2013)9009.
108. T. Aaltonen et al., Phys. Rev. Lett. 99, 151,801 (2007).
109. ATLAS and CMS Collaborations, Phys. Rev. Lett. 114 (2015) 191803 and arXiv: 1503.07589.
110. V.Khachatryan et al. JINST 10 (2015) P06005 and arXiv 1502.02701.
111. ATLAS Detector and Physics Performance Technical Design Report, CERN-LHCC/99–14.
112. N.Cartiglia et al. NIM A850 (2017)83 and arXiv:1609.08681
113. ATLAS Collaboration (2017) arXiv 1708.04445.
114. R. Leitner et al., ATL-TILECAL-PUB-2007-002.
115. G. Aad et al., ATL-COM-PHYS-2008-243, Performance Chapter, p. 41.
116. ATLAS Muon Spectrometer Technical Design Report (1997) CERN/LHCC/97–22.
117. M. Bosman, ATL-CONF-2002-001.
118. S. Catani et al., Phys.Lett. B 285(1992)291.
119. M. Cacciari, G. Salam, G. Soyez, JHEP 0804:063 (2008).
120. G. Aad et al., ATLAS Collaboration, J. Instrum. 3 (2008) S08003.
121. S. Chatrchyan et al., CMS Collaboration, J. Instrum. 3 (2008) S08004.
122. CMS Collaboration, JINST 12 (2017) P10003 and arXiv 1706.04965.
123. CMS Collaboration, JINST 12 P02014 (2017) and arXiv 1607.03663.
124. ATLAS Collaboration, Phys.Rev. D96 072002 (2017) and arXiv 1703.09665.
125. CMS Collaboration, JINST 6 P09001 (2011) and arXiv:1106.5048.
126. ATLAS Collaboration, Eur. Phys. J. C77 (2017) 241.
127. ATLAS Collaboration, Eur. Phys. J. C77 (2017) 317 and arXiv:1611.09661.
128. T.Head for the LHCb Collaboration JINST 9 (2014) C09015.
129. R. Hofstadter, Phys. Rev. 74 (1948) 100.
130. R.R. Wilson, Nucl. Instrum. Meth. 1 (1957) 101.
131. J. Engler et al., Nucl. Instrum. Meth. 106 (1973) 189.
132. H. Schopper, Proc. Int. Conf. History of Original Ideas and Basic Discoveries in Particle Physics, Erice, 1994.
133. A.M. Baldini et al., Eur. Phys. J. C76 (2016) 434 and arXiv: 1605.05081.
134. M. Baldini et al. (MEG II Collaboration), PSI Proposal R-99-05.2) (2013) arXiv:1301.7225.
135. E.Aprile et al. Eur. Phys. J. C77 (2017) 881 and arXiv: 1705.01828.
136. M. Aharrouche et al., Nucl. Instrum. Meth. A 568 (2006) 601.
137. ATLAS Collaboration. Eur. Phys. J. C. 78(2018) 110, arxiv 1701.07240.
138. M. Derrick et al., Nucl. Instrum. Meth. A 309 (1991), 77.
139. B. Andrieu et al., Nucl. Instrum. Meth. A 336 (1993) 460.
140. T. Akesson et al., Nucl. Instrum. Meth. A 241(1985) 17.
141. ATLAS Collaboration, CERN/LHCC/96–042 (1996).
142. CMS Collaboration, CERN/ LHCC 97–31 (1997).
143. L. Linsen et al., CERN Yellow report CERN-2012-003 (2012).
144. N. Akchurin et al., Nucl. Instrum. Meth. A 584(2008)273.
145. A. Abada et al., FCC physics opportunities, Future Circular Collider Conceptual Design Report Vol. 1, Eur. Phys. J. C. 79, 474 (2019). https://doi.org/10.1140/epjc/s0052-019-6904-3

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 7
Particle Identification: Time-of-Flight, Cherenkov and Transition Radiation Detectors

Roger Forty and Olav Ullaland

7.1 Introduction

Particle identification, PID, is of crucial importance in most experiments. The requirement can range from positive π/K identification in B-physics channels like $B_s^0 \rightarrow D_s^{\mp} K^{\pm}$ against a background from $B_s^0 \rightarrow D_s^- \pi^+$ which is ~15 times more abundant, to e/π separation at the level of $\sim 10^{-2}$ for momenta >1 GeV/c in order to effectively suppress a combinatorial background in channels like leptonic decays of heavy vector resonances.

The detectors should be non-destructive and should in addition introduce as little radiation length or interaction length as possible. We will in this chapter examine three experimental techniques which can be deployed for charged particle identification.

That is Time-of-Flight, Sect. 7.2, and Cherenkov detectors which measure the particle velocity relative to the speed of light in vacuum, $\beta = v/c$, Sect. 7.4, and transition radiation detectors which are sensitive to $\gamma = 1/\sqrt{(1-\beta^2)}$ of the charged particle, Sect. 7.5. These detectors cannot be stand-alone detectors for PID purposes. They all require that the momentum of the particle is defined by other means, see Sect. 4.3, and then

$$\frac{m^2}{p^2} = \frac{1}{\gamma^2 - 1} = \frac{1-\beta^2}{\beta^2} \tag{7.1}$$

allowing the mass m of the particle to be determined.
Only a limited amount of theory is included in this chapter as this is covered elsewhere in this book. The main emphasis will be on the working principles of these

R. Forty (✉) · O. Ullaland
CERN, Geneva, Switzerland
e-mail: Roger.Forty@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_7

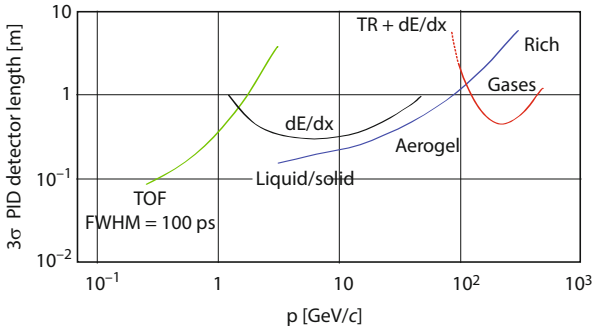


Fig. 7.1 Pion-kaon separation by different PID methods: the length of the detectors needed for 3 sigma separation. Adapted from [1]

detectors and how they are incorporated into compound experiments. A graphic representation of the different identification techniques can be seen in Fig. 7.1.

7.2 Time of Flight Measurements

The mass identification, m_i, of a momentum defined, p_i, charged particle is straight forward by measuring the flight time, t_i, over a path length, l. The mass, momentum, path length and flight time are related by:

$$m_i^2 = \frac{p_i^2}{l^2}\left[ct_i - l\right]\left[ct_i + l\right] \tag{7.2}$$

and the uncertainties by:

$$\left(\frac{\Delta m}{m}\right)^2 = \left(\frac{\Delta p}{p}\right)^2 + \gamma^4\left[\left(\frac{\Delta t}{t}\right)^2 + \left(\frac{\Delta l}{l}\right)^2\right] \tag{7.3}$$

There are essentially two sources of errors[1] in the measurement of time, t, in Eq. (7.3).

1. The limitation of the electronics to resolve short time intervals. A random time jitter in the pulse height at the detector and thereby a time slewing or time walk.
2. Variation of the transit time of the photons or the free electrons, and thereby the signal formation time, in the detector.

[1] *Irresolution* was proposed in [2]. Although a nice word, it did not catch on.

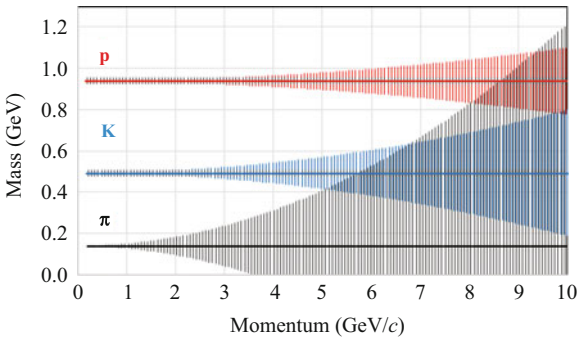


Fig. 7.2 Mass resolution as function of momentum for a Time of Flight, ToF, detector with $\Delta p/p = 4 \cdot 10^{-3}$, $l = 10\,\text{m}$, $\Delta l/l = 10^{-4}$ and $\Delta t = 50\,\text{ps}$

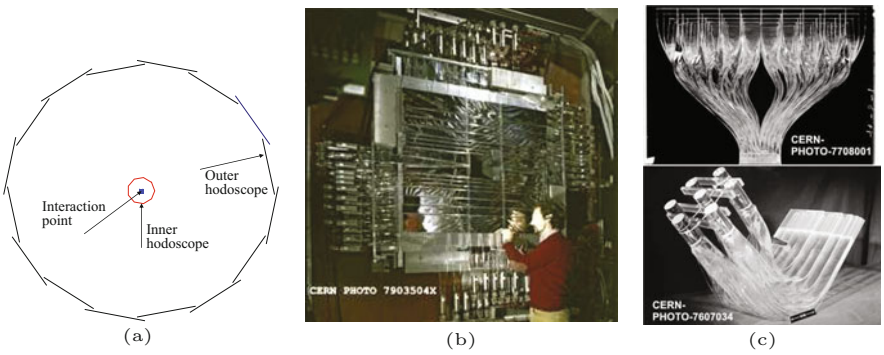


Fig. 7.3 (**a**) Simplistic sketch of a Time of Flight system. (**b**) Large scintillator hodoscope from CERN experiment NA1. (**c**) Light guides and scintillators

Particle misidentification will therefore occur when the time difference between two particles with the same momentum becomes comparable to the detector resolution. Figure 7.2 gives the mass resolution as function of momentum for π, K and proton.

Time of Flight detectors, ToF, have throughout been essential tools in physics experiments and have undergone impressive improvements in time resolution from micro-seconds to pico-seconds. The basis was worked out in [3]. A principle sketch is given in Fig. 7.3a in the Centre of Mass coordinate system. The interaction point is surrounded by a *time zero* hodoscope, the Inner hodoscope. Another hodoscope, the Outer hodoscope, is placed at a distance l from the first one. Assuming that there is a momentum measurement between the two, this is all that is needed to solve Eq. (7.2).

The Inner hodoscope is usually not required. In a colliding beam experiment, the RF structure can be adequate to give a sufficiently precise *time zero*. In events with

a large number of secondaries, one can use the feature that at least one particle will have a velocity $v \cong c$ and thereby use this one to define *time zero*.

The main work during the last years [4] has been in the improvement of the time resolution and, as the detectors have gradually increased in size, in the cost/m^2. The occupancy and radiation tolerance are playing a very important role for detectors that are proposed for the new high luminosity accelerators. We will here not explain the working principle of the detectors themselves. The reader is referred to Chap. 3. We will rather discuss the advantages and inconveniences of some of the most commonly used detector set-ups.

7.2.1 Scintillator Hodoscopes

A scintillator, read out in both ends by a photomultiplier, is the classic element of a Time of Flight hodoscope, Fig. 7.3b. The number of photons created is large. Plastic scintillators, as discussed in Chap. 3, have a density $\rho \simeq 1.03\,\mathrm{g/cm^3}$. About 10^4 photons/MeV are created with a mean wavelength of ∼400 nm and a time constant $\tau \sim 1.5\,\mathrm{ns}$. The number of emitted photons per time unit, N, will be approximately:

$$N = \frac{N_0}{\tau} \exp\left(-\frac{t}{\tau}\right) \tag{7.4}$$

N is the number of photons emitted at time t, N_0 is the total number of emitted photons and τ is the average lifetime. τ is characteristic to a specific scintillator material. A short decay time increases the maximum count rate and is therefore an important property for detection. Most inorganic scintillators have rather long decay times, $\tau \sim 100\,\mathrm{ns}$, but in some cases the decay constant can be very short. For example, $\tau = 1\,\mathrm{ns}$ for BaF_2.

The specific energy loss, Chap. 2, for a minimum ionizing particle, MIP, is given as:

$$\left(-\frac{dE}{dx}\right)_{\min} = 2.35 - 1.47\ln(Z) \quad \mathrm{MeV cm^2/g} \tag{7.5}$$

where Z is the atomic number.

$(dE/dx)_{\min}$ for a plastic scintillator is about $2\,\mathrm{MeV\,cm^2/g}$, or about $2 \cdot 10^4$ photons/cm are produced. This number of detectable photons will be greatly reduced due to the attenuation length of the material, the losses out from the material, quantum efficiency of the photon detector and the shaping time of the electronics. As the final number of photoelectrons is heavily dependent on the exact lay-out of the detector, it is very difficult to give a *typical* number. But, as a rule of thumb, approximately $2 \cdot 10^{-3}$ photoelectrons will be produced by the primary photon. This would give in the range of 40 photoelectrons/cm in a plastic

scintillator. Let N_D be the total number of detected photons. The time resolution is roughly proportional to $1/\sqrt{N_D}$. ToF detectors with high resolution, $\Delta t \sim 100$ ps, therefore use scintillator thickness of 2–3 cm. The material budget then becomes important.

The connection between the scintillator and the photon detector is a very important step in order to maximise the light collection efficiency of the system. These light concentrators are normally built around a Winston Cone [5] or a fishtail as in Fig. 7.3b. A Winston Cone is a non-imaging off-axis parabola of revolution which will maximise the collection of incoming rays. The ideal concentrator will achieve the highest possible concentration of radiant energy permitted by the second law of thermodynamics. This is equivalent to the general theorem of Liouville [6]. More specific in a case of a light guide, one can write:

$$n^2 - 1 \geq \left[\frac{d}{2r} + 1\right]^2 \tag{7.6}$$

where d is the light guide diameter and r is the bending radius. n is the refractive index relative to air. See Fig. 7.3c. Charged particles going through the light guides will give signal due to Cherenkov radiation and thereby give rise to an event correlated background.

A well designed scintillator for ToF must provide a good photon collection efficiency and a small time jitter. For fast timing one would normally rely on the first direct photon impact. This puts further constrains on the photon detector. The classic photon detector is the photomultiplier tube (PMT). Depending on the window geometry, dynode chain and HV configuration, the transient time spread is in the range of 1 ns. This can be reduced by instrumenting both ends of the scintillator and then use mean timing. This will also take care of the after-pulsing in the PMT. These are normally either due to ions in the residual gas in the PMT which drift back, strike the photo cathode and liberate new photoelectrons or light from the dynodes which hit the photo cathode. The first will give a signal about 100 ns after the event, while the latter signal comes after 30–60 ns. See Chap. 3 for more information. However, still to overcome the path length and transient time variation, the detector has to output a large amount of primary photons to achieve total time resolution in the range of 100 ps.

An example can be found in [7]. Mean timing and time slewing corrections are performed. Slew-correction time, T^{cor}, is defined as:

$$T^{\text{cor}} = T + \frac{A_0}{\sqrt{\text{ADC}}} \tag{7.7}$$

where the constant A_0 is normally evaluated for each PMT and ADC is the signal pulse height. They report a nearly constant time resolution of $\sigma \sim 55$ ps across a detector length of 15 cm.

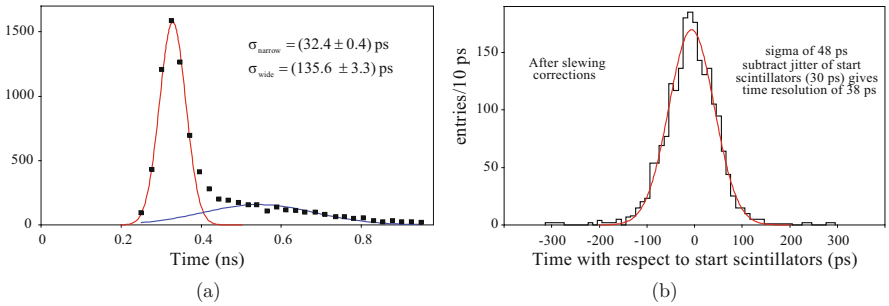


Fig. 7.4 (**a**) Single photoelectron timing resolution in Burle 64-pixel MCP-PMT 85012-501 with 10 μm hole diameter. Adapted from [10]. (**b**) Time distribution of MRPC after slewing corrections. Adapted from [11]

Other photon detectors are generally faster and with smaller time spread than the PMT. See Chap. 3 for a detailed description of these devices. Below are some listed from [8]:

- 100 μm diameter GaP SiPMT Avalanche Photo Diode operating in a Geiger mode with active quenching [9]. Single photoelectron regime: 25 ps
- Hamamatsu H-8500 Flat panel multi anode photo multiplier tube (MaPMT).[2] SLAC measurement [8] of single photon resolution: 140 ps
- Burle 85011 photo multiplier tube with micro channel plate (MCP-PMT).[3] SLAC measurement [8] of single photon resolution: <50 ps

A drawback with these detectors can be the non-Gaussian tails as shown on Fig. 7.4a.

7.2.2 Parallel Plate ToF Detectors

One of the main challenges in using gas based detectors, MWPC up to spark chambers as discussed in Chap. 3, is the time jitter caused by the spread in pulse heights due to the long Landau tail. This can to some extent be overcome by using many gaps and operating the detector in a regime where the pulse height is nearly independent of the primary ionisation. However, this can seriously diminish the rate capability of these detectors. Well adapted electronics will furthermore decrease the time walk.

[2] HAMAMATSU PHOTONICS K.K. 325-6, Sunayama-cho, Naka-ku, Hamamatsu City, Shizuoka Pref., 430-8587, Japan.

[3] BURLE INDUSTRIES, INC. 1000 New Holland Avenue, Lancaster, PA 17601-5688 U.S.A.

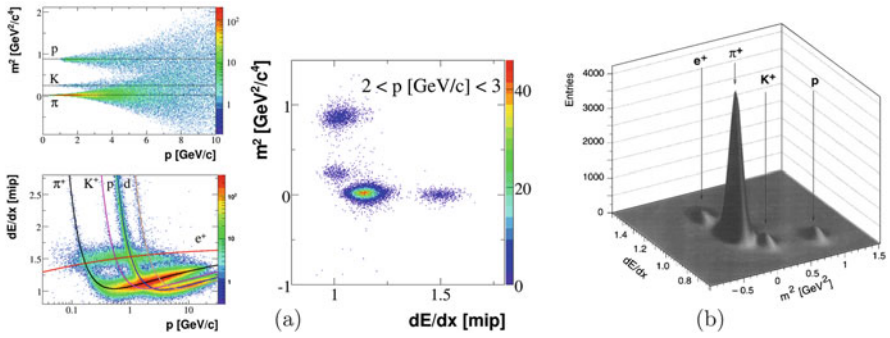


Fig. 7.5 (**a**) Particle identification in NA61. Reference [13]. (**b**) Particle identification at NA 49 by simultaneous dE/dx and TOF measurement in the momentum range 5 to 6 GeV/c for central Pb+Pb collisions. Reference [14]

Large area resistive plate chambers, see Chap. 3, are successfully used as time of flight detectors. An example is the $\sim$150 m^2, with $1.6 \cdot 10^5$ read-out channels, detector for ALICE [11]. Ten gaps of 250 μm width are made from 400 μm thick soda-lime glass with a gas composition of $C_2H_2F_4$:SF_6:C_4H_{10} = 0.90:0.05:0.05. The resistivity[4] of the glass is $\sim 10^{13}$ Ωcm. The detector is operated just below streamer mode. Tests indicate no change in performance up to 1 kHz/cm^2. As there are many gaps, the output charge distribution is a broad, but nearly Gaussian distributed with some Landau tail towards higher value. This will give rise to some time slewing. The time resolution is given as $\sigma < 40$ ps. See Fig. 7.4b.

7.3 The Power of Combined PID

The inherently simple ToF technique has greatly evolved over the years. The coming of the higher energy and/or higher intensity accelerators have required an ever better time and space resolution. Even though there has been great progress with small single pixel devices, progress with large systems has been slow. An overview of the current state of the art can be found in [12].

Combining different PID techniques, even with modest resolution, has been the preferred option for many experiments. An example of this powerful approach is shown in Fig. 7.5.

[4] It can be worth noting that materials which exhibit very large resistivity, might not be Ohmic, but rather ionic, and thereby show large variations depending on the applied current or voltage.

7.4 Cherenkov Radiation

The theory of Cherenkov radiation is discussed in Chap. 2. Further reading can be found in references [15–18]. We will here just recall some of the main features. The condition for emission of a Cherenkov photon is given by

$$\cos\Theta_C = \frac{1}{\beta \cdot \sqrt{\boldsymbol{\varepsilon(\lambda)}}} = \frac{1}{\beta \cdot n(\lambda)} \tag{7.8}$$

and the number of emitted photons by

$$\frac{d^2N}{dL\,d\lambda} = 2\pi\alpha Z^2 \frac{\sin^2\Theta_C}{\lambda^2}, \tag{7.9}$$

where Θ_C is the angle of the emitted photon relative to the particle trajectory, ε is the dielectric constant as function of the photon wavelength λ, L is the radiator length, $\alpha \sim 1/137$ is the fine structure constant, β is the particle velocity relative to the speed of light in vacuum, $\beta = v/c = pc/E$, and Z is the charge of the particle in units of electron charges. The refractive index, n, is given as $n^2 = \varepsilon$. The relationship between the photon wavelength and its angular frequency, ω, is given by $\lambda(\text{nm}) \sim 1240/\hbar\omega(\text{eV})$. A representation of the Cherenkov radiation domain is given in Fig. 7.6a.

From the discussion in Chap. 2 and Eq. (7.8) it is clear that ε has to be real and larger than 1 and that the speed of the charged particle must be larger than the phase velocity of the electromagnetic fields at the frequency ω in order to have emission of Cherenkov photons at that frequency.

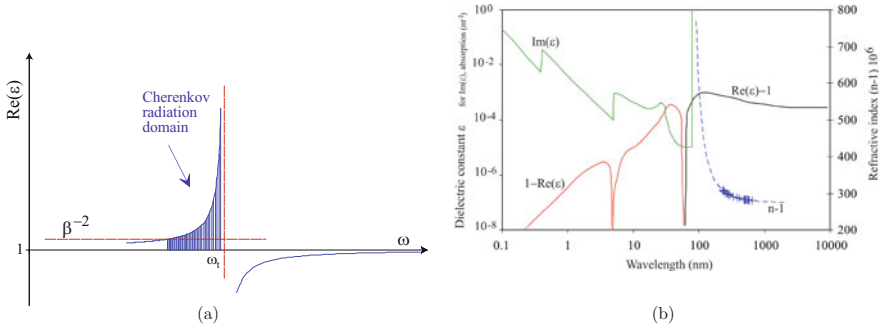


Fig. 7.6 (**a**) Simplistic representation of the real part of the dielectric constant, $\Re(\varepsilon)$, as function of the frequency, ω. (**b**) The dielectric constant, ε, and the refractive index, n, for argon at 0 °C and 101.3 kPa. ε data replotted from [19] and n from [20]

We see from the above that Cherenkov radiation is characterized by:

- Cherenkov radiation is a prompt signal.
- The existence of a threshold[5] in $\beta_{\min} = n^{-1}$
- The Cherenkov angle is depending on β.
- The number of Cherenkov photons emitted is depending on β.
- The number of photons emitted is depending on the square of the charge of the particle.

The properties described above of Cherenkov radiation can be used to measure the velocity of a charged particle traversing matter. Consider two charged particles with known momenta p and mass and velocity given by m_i and β_i. The mass difference can then be written as:

$$m_1^2 - m_2^2 = p^2 \cdot \frac{(\beta_1 - \beta_2)(\beta_1 + \beta_2)}{(\beta_1 \cdot \beta_2)^2} = n^2 p^2 \cdot (\cos^2 \Theta_1 - \cos^2 \Theta_2) \tag{7.10}$$

And if $n - 1$ is small

$$m_1^2 - m_2^2 = p^2 \cdot (\Theta_2 + \Theta_1)(\Theta_2 - \Theta_1) \tag{7.11}$$

The resolution in mass is thereby directly linked to the angular resolution of the detector. The main emphasis for all the Cherenkov detectors will be angular resolution.

The refractive index together with ε, for argon at 0 °C and 101.3 kPa, is given in Fig. 7.6b. The data for the refractive index of argon is well described by a single pole Sellmeier, see Eq. (7.16), representation:

$$(n - 1) \cdot 10^6 = \frac{0.05086}{73.82^{-2} - \lambda^{-2}} \tag{7.12}$$

with λ in nm. We observe that this pole is where $\Re(\varepsilon)$ goes from larger than 1 to smaller than 1. At about the same wavelength $\Im(\varepsilon)$ becomes important.

A Cherenkov light detector is therefore based on classical optics. The choice of radiator, and thereby the refractive index, is depending on the momentum range which has to be covered and the photon detector option. We will in the following discuss different radiator materials, Sect. 7.4.2, and the usage from Threshold, Sect. 7.4.3, to Ring Imaging Cherenkov detectors, Sect. 7.4.4. We will first take a closer look at the refractive index, Sect. 7.4.1.

[5] Due to diffraction broadening, Cherenkov photons can be emitted below threshold. We will not discuss that here.

Table 7.1 Atomic refraction constants from Ref. [22]

Atom		Atomic refraction
Carbon		2.418
Bromine		8.865
Chlorine		5.967
Fluorine		1.1
Hydrogen		1.1
Iodine		13.952
One double bond	=O	2.122
Two single bonds	−O−	1.643

7.4.1 Refractive Index

The dielectric constant is given by:

$$\varepsilon = 1 + 4\pi\chi = \frac{1 + \frac{8}{3}\pi \mathrm{N}\zeta}{1 - \frac{4}{3}\pi \mathrm{N}\zeta} \quad \text{from which} \quad \frac{4}{3}\pi \mathrm{N}\zeta = \frac{\varepsilon - 1}{\varepsilon + 2} \tag{7.13}$$

where χ is the susceptibility, N is the number of molecules per unit volume and ζ is the molecular polarizability.

A relation like this was first obtained by Mossotti in 1850, then by Lorenz in 1869, and refined by Clausius in 1879, and which is usually called the Clausius-Mossotti equation. Polarizable matter was modelled as an assembly of small conducting spheres in the early studies.[6] Maxwell's theory showed that the index of refraction of light, n, was related to ε by $n^2 = \varepsilon$, so that the formula could be applied to light as well as to static fields. H.A. Lorentz, in 1878, and L.V. Lorenz (1829–1891), in 1881, derived a similar formula on the basis of the electron theory in which n^2 replaced ε. This formula is called the Lorentz-Lorenz formula, and can be written in the following way:

$$n^2 = \frac{1 + 2\left(\rho \frac{R_M}{M_W}\right)}{1 - \left(\rho \frac{R_M}{M_W}\right)} \tag{7.14}$$

where R_M is the molar refraction, M_W is the molecular weight and ρ is the density.

The molar refraction may then be estimated from the chemical formula. Atomic refraction constants differ slightly in the literature, but the constants in Table 7.1 give reasonable results for many compounds.

The Lorentz-Lorenz equation, Eq. (7.14) together with Table 7.1, does not explicitly express the refractive index as a function of the photon energy.

[6] Strictly speaking, Clausius-Mossotti equation is only rigorously valid in the limit of zero density [21].

The most common way to represent the refractive index is in the form of a series with multiple poles

$$n-1=C\cdot\sum_i\frac{f_i}{\nu_i^2-\nu^2}\quad\text{with}\quad C=\frac{e^2A}{2\pi mc^2}=1.2098\cdot10^6 \tag{7.15}$$

where e and m are the charge and mass of the electron, A is Avogadro's number per cm^3 and $\nu(\mathrm{cm}^{-1})=10^7/\lambda(\mathrm{nm})$. f_i is the oscillator strength of the Eigen frequency ν. We will here mainly use the standard Sellmeier formula with one pole:

$$\frac{3}{2}\cdot\frac{n^2-1}{n^2+2}=\frac{a}{\lambda_0^{-2}-\lambda^{-2}}\simeq n-1\quad\text{for}\quad n-1\ll1 \tag{7.16}$$

$b=\lambda_0^{-2}$ will also be used. λ is in nm. a/b is the asymptotic value of n as $\lambda\to\infty$. A two pole Sellmeier representation might be required:

$$\frac{3}{2}\cdot\frac{n^2-1}{n^2+2}=\left[a_1\cdot\lambda^{-4}+a_2\cdot\lambda^{-2}+a_3\right]^{-1} \tag{7.17}$$

Clearly also other types of power series can be used to approximate the refractive index like in reference [23]. In this case the refractive index is approximated with the half empirical formula of a n-term Cauchy equation which is very similar to Eq. (7.17):

$$n-1=2\pi N_0\left[a_0+a_1\omega^2+a_2\omega^4+a_3\omega^6\right] \tag{7.18}$$

where ω is the frequency in atomic units. The $a_3\omega^6$ term has been added after the original series [24] was truncated at $a_2\omega^4$ and thereby was not very useful in the UV to VUV wavelength region.

The refractive index of a medium M which is a mixture of different molecules in the ratio $M=\sum_i[M_i/f_i]$ for $1=\sum_i f_i^{-1}$, is given by $n_M=\sum_i[n_i/f_i]$. We will illustrate this with a simple example. The refractive index of air and its constituents are well measured quantities, Fig. 7.7a.

The Sellmeier parameterisation for N_2, O_2, CO_2 and argon is given in Table 7.2. Note that whereas a single pole, Eq. (7.16), describes well N_2, CO_2 and argon, the data for O_2 is best described with a two pole, Eq. (7.17), representation. The parameters used to describe the data points for dry air in Fig. 7.7a are

$$\begin{aligned}n(\mathrm{air})&=0.7809\cdot n(\mathrm{N_2})+0.2095\cdot n(\mathrm{O_2})+0.0093\cdot n(\mathrm{Ar})+0.0003\cdot n(\mathrm{CO_2})\\&\simeq1+10^{-6}\cdot\left[\left(\lambda^{-2}-69.1^{-2}\right)\left(\lambda^{-2}+99.5^{-2}\right)\right]^{-1}\end{aligned} \tag{7.19}$$

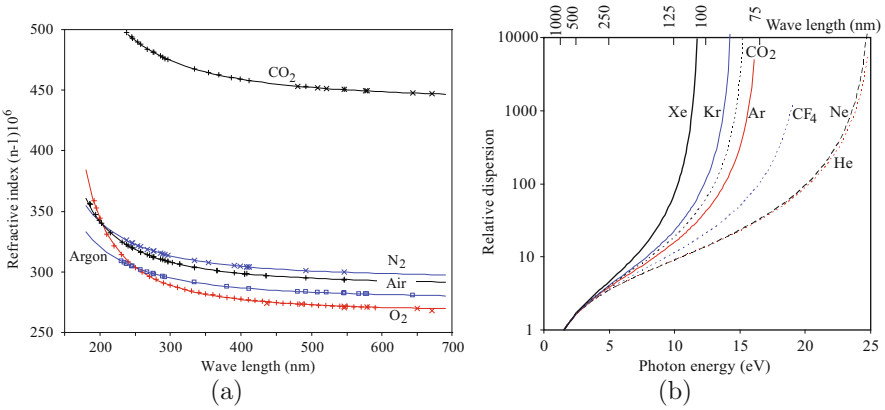


Fig. 7.7 (**a**) The refractive index of dry air, N_2, O_2, CO_2 and argon at 0 °C and 101.3 kPa [20]. (b) Dispersion dn/dE relative to the value at 800 nm, in some noble and n-atomic gases as function of the photon energy [20]

Table 7.2 Sellmeier fit, Eqs. (7.16) and (7.17), parameters for the gases at 0 °C and 101.3 kPa

Gas	A	B	a_1	a_2	a_3	λ_0
N_2	0.0532	0.000181				74.36
O_2			−54,955	−20.275	0.00376	122.90
CO_2	0.0687	0.000156				80.10
Ar	0.0509	0.000184				73.82

The pole, λ_0, in nm. O_2 has only one real pole in this representation

Although the last expression gives a good description of the refractive index for dry air at 0 °C, 101.3 kPa and for $\lambda \geq 130$ nm, the real pole at ∼69 nm has no physical meaning.

7.4.2 Cherenkov Radiators

Cherenkov radiators have to be reasonably optically transparent and with an appropriate refractive index. The scintillation and phosphorescence processes in the medium should be small. There is a wide variety to chose from, from transparent solids via liquids to gases. One can in addition change the refractive index by changing temperature and pressure of the medium.

The dispersion in a radiator can be written as

$$\frac{dn}{dE} \propto \frac{(n^2-1)^2}{n} \cdot E \quad \text{and for } (n-1) \ll 1 \text{it reduces to} \quad \frac{dn}{dE} \propto (n-1)^2 \cdot E \tag{7.20}$$

where E is the energy of the photon.

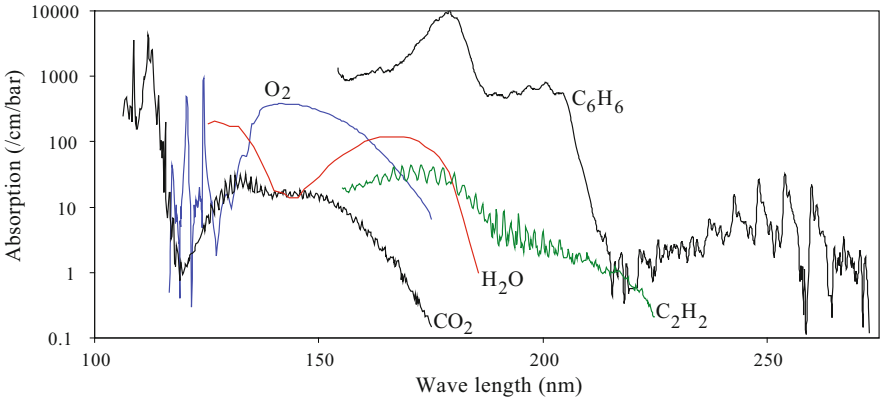


Fig. 7.8 Absorption as function of the photon wavelength [25]

The dispersion in some noble and n-atomic gases is plotted in Fig. 7.7b. He and Ne are very weakly dispersive in contrast to Kr and Xe. As can be seen from Fig. 7.7b, fluorocarbons are also weakly dispersive. If the definition of the Cherenkov angle is an important quantity for the detector, it is then clear that the dispersion has to be as small as possible over the photon detector efficiency window. The detector design will be a balance between number of photons and the spread in the Cherenkov angle.

The radiator medium becomes opaque when the imaginary part of the dielectric constant becomes important, Fig. 7.6b. Most media will in addition exhibit broad and strong absorption bands. Figure 7.8 shows the absorption in some commonly used Cherenkov media or trace impurities in them. For simple alkanes, C_NH_{2+2N}, the onset of photon absorption [25] can be approximated to:

$$\lambda_{CH}(\text{nm}) = 181 - \frac{226}{3(N+1)} \tag{7.21}$$

A similar approximation can be given for n-perfluorocarbons, C_NF_{2+2N},:

$$\lambda_{CF}(\text{nm}) = 175.6 - \frac{641}{3N+5.7} \tag{7.22}$$

It can be seen from these two expressions that n-perfluorocarbons are more transparent than alkanes. Alkanes are therefore good quenchers as used in MWPCs. Trace impurities are particularly difficult to eliminate especially when it is not clear which molecule is causing the absorption. The successful detector design should therefore not be sensitive to these bands.

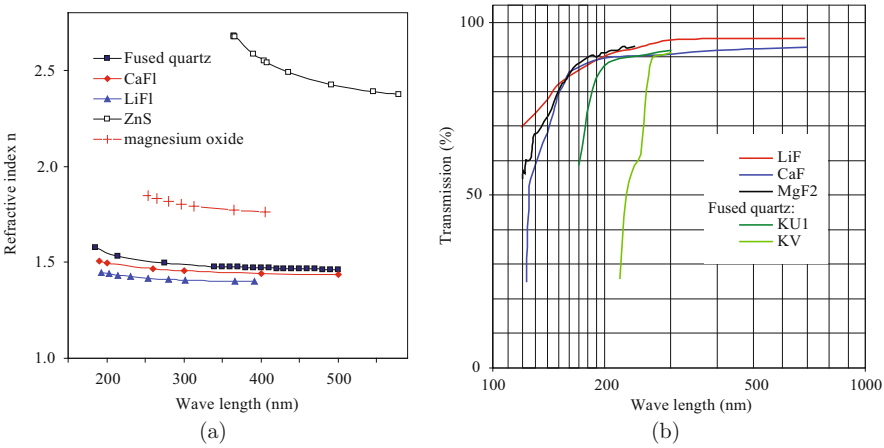


Fig. 7.9 (**a**) Refractive index for quartz and other special optical materials [26]. (**b**) Transmission in some commercially available quartz as function of wavelength. See footnote 7

7.4.2.1 Quartz Radiators

Quartz radiators are very popular for Cherenkov detectors operating in the low momentum range. The refractive index for some quartz and other optical materials is given in Fig. 7.9a. Figure 7.9b gives the transmission for some commercially[7] available quartz. By choosing a refractive index $n \sim 1.5$ and a photon detection window from 800 to 300 nm, the Cherenkov angle measurement between π and K becomes difficult for $p > 2\,\text{GeV}/c$ due to dispersion.

Quartz radiators in Cherenkov detectors are treated in two distinctly different ways. We see that for the $n \sim 1.5$ quartz, a π will pass the Cherenkov threshold at 125 MeV/c and at ∼280 MeV/c no light will escape the quartz due to total internal reflection for particles perpendicular onto the radiator. An elegant solution to this problem is shown in Fig. 7.10a.

The other option is to exploit the feature of internal reflections as a light guide for the Cherenkov photons. The working principle of a DIRC, Detection (of) Internally Reflected Cherenkov (light) [28], detector is sketched in Fig. 7.10b. The standoff region is designed to maximize the transfer efficiency between the radiator and the detector. If this region has the same index of refraction as the radiator, $n_1 \simeq n_2$, the transfer efficiency is maximized and the image will emerge without reflection or refraction at the end surface. Further improvements can be achieved by measuring the transfer time of the Cherenkov photons [29]. A large fraction of the uncertainties caused by the dispersion can then be eliminated.

[7] Data from:
Del Mar Ventures, 12595 Ruette Alliante No.148, San Diego, California 92130, US.
Crystran Ltd, 1 Broom Road Business Park, Poole, Dorset, England.

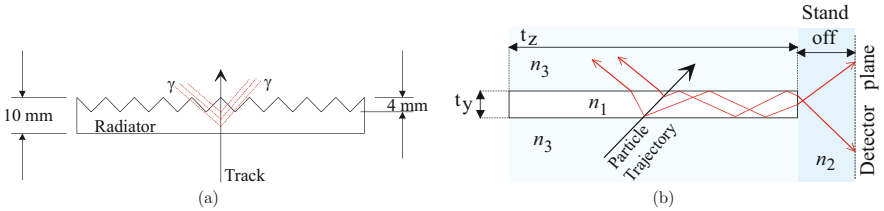


Fig. 7.10 (**a**) Sketch of the saw tooth quartz radiator for CLEO 3 [27]. (**b**) Schematic of the radiator bar for a DIRC [28] detector. Not to scale

Similar, but not identical, are the Time-of-Propagation, TOP [30] detector at the BELLE II experiment and the proposed detectors; TORCH [32] at LHCb and a DIRC [33] at the PANDA experiment.

The TOP consists of quartz radiator bars 270 cm long $\times$45 cm wide $\times$2 cm thick. See Fig. 7.11a. One end of the bar has a spherical mirror to reflect light back to the other end that has a small expansion prism. The prism is instrumented with 32 16-channel microchannel plate photomultiplier tubes (MCP-PMTs) readout with custom giga-sample per second waveform sampling electronics. The Cherenkov ring is imaged by the 512 MCP-PMT pixels with 5 mm pixel size and the time of arrival of each photon is measured with $<$50 ps timing resolution. The photon time of arrival is a sum of the time of flight of the charged particle to the quartz radiator and the time of propagation of the Cherenkov photons to the photodetectors. Results with test beam data is shown in Fig. 7.11b. Clear π/K separation can be observed. The detector will be ready for data taking in 2018.

7.4.2.2 Aerogel Radiators

The search for a stable Cherenkov radiator with a refractive index between gas and liquid started about the same time as the first Cherenkov detector became operational. The first successful was silica aerogel [34]. The Axial Field Spectrometer [35] at the CERN ISR was the first large experiment to use it. The principle fabrication reactions[8] are rather simple:

$$\begin{aligned} Si(OCH_3)_4 + 4H_2O &\xrightarrow{NH_3} Si(OH)_4 + 4CH_3OH \\ nSi(OH)_4 &\longrightarrow (SiO_2)_n + 2nH_2O \end{aligned}$$

The refractive index, n, as a function of the wavelength, λ, can be approximated by $n = 1 + k \cdot \rho$ for the density ρ in g/cm^3 and k is a function of λ. An example

[8] Tetramethyl orthosilicate is used in this example. Tetraethyl orthosilicate can also be used and is normally preferred as the byproduct is ethanol rather than methanol. Both tetramethyl orthosilicate and tetraethyl orthosilicate are highly reactive [36].

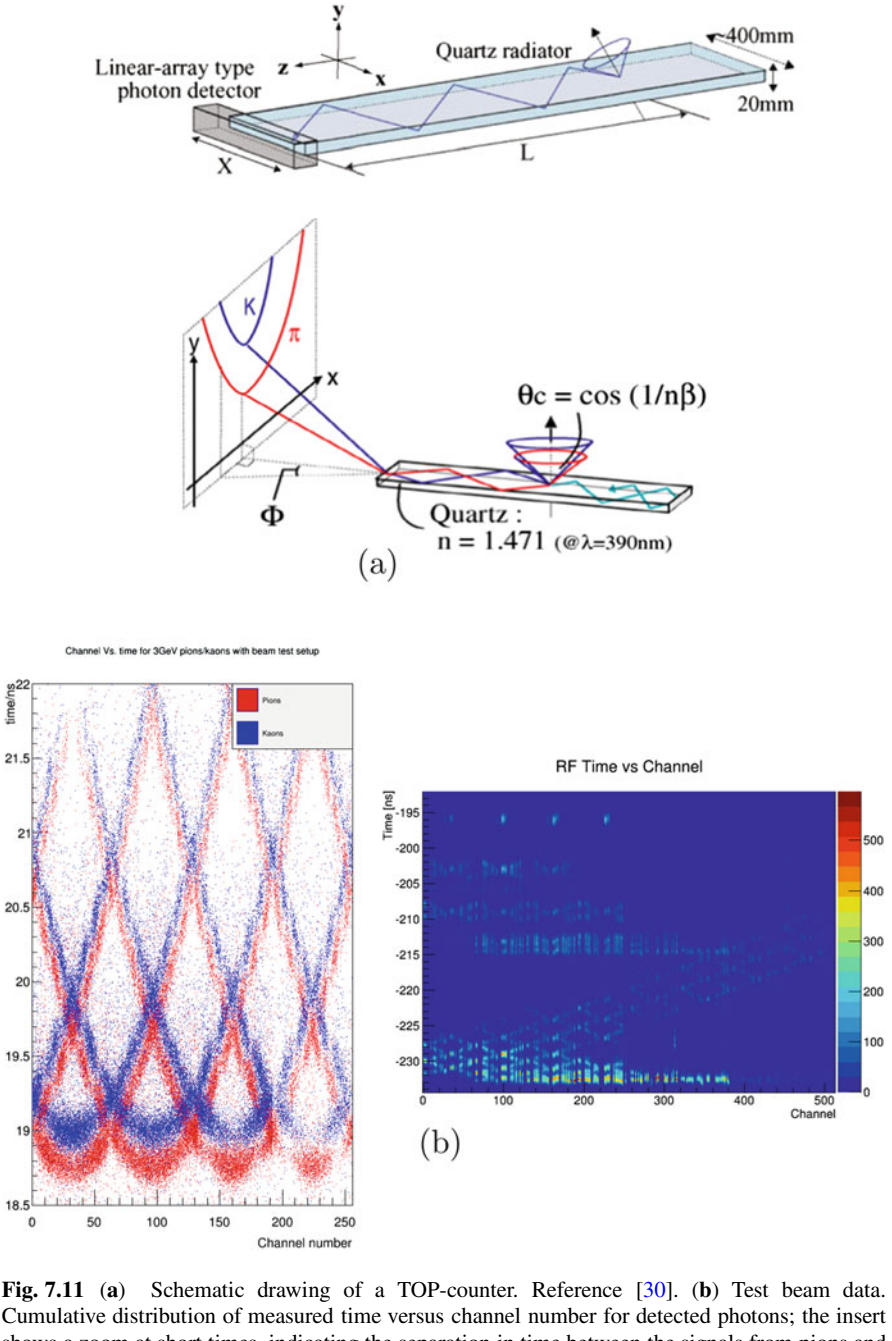


Fig. 7.11 (**a**) Schematic drawing of a TOP-counter. Reference [30]. (**b**) Test beam data. Cumulative distribution of measured time versus channel number for detected photons; the insert shows a zoom at short times, indicating the separation in time between the signals from pions and kaons. Reference [31]

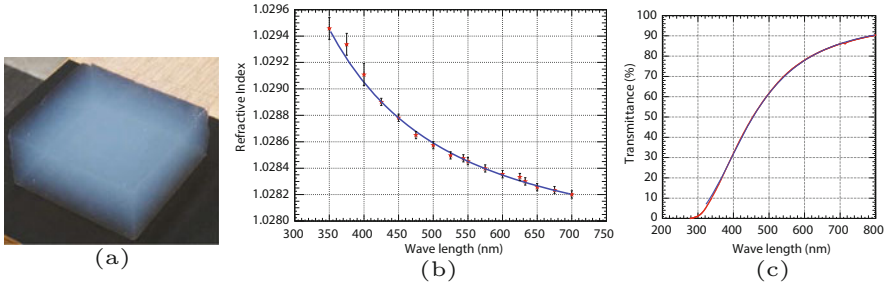


Fig. 7.12 (**a**) Aerogel tile. Courtesy of the LHCb Milano group. (**b**) Refractive index of aerogel as function of wavelength. Bellunato et al. [37] with permission. (**c**) Transmittance of 52.10 mm thick aerogel as function of wavelength. Perego [38] with permission

is shown in Fig. 7.12b. The data [37] is well described by a single pole Sellmeier equation:

$$n^2 - 1 = \frac{a_0 \lambda^2}{\lambda^2 - \lambda_0^2} \tag{7.23}$$

for $a_0 = 0.05639 \pm 0.00004$ and $\lambda_0 = (83.22 \pm 1.25)$ nm.

Assuming that aerogel is just a rarefied form of silica, a_0 and the density of the material are linked by:

$$\rho(\text{aerogel}) = \frac{a_0(\text{aerogel})}{a_0(\text{SiO}_2)} \frac{n^2(\text{SiO}_2) + 2}{n^2(\text{aerogel}) + 2} \rho(\text{SiO}_2) \tag{7.24}$$

which gives $\rho(\text{aerogel}) = (0.158 \pm 0.001)\,\text{g/cm}^3$, in reasonably good agreement with $\rho = (0.149 \pm 0.004)\,\text{g/cm}^3$ which was given by the manufacturer.

Two main types of aerogel are now available, hydrophobic[9] and hygroscopic.[10] Large homogeneous blocks of high optical quality are now readily available. The refractive index can be tuned between 1.008 and 1.1. By stacking aerogel blocks of different refractive indices, the total light output can be increased while minimizing the width of the Cherenkov ring. By modifying the reaction conditions of the sol-gel synthesis [39], it is possible to control the variations of n inside the aerogel tile and thereby create a monolithic block with well defined different layers of n.

[9] Advanced Technology Research Laboratory, 1048 Kadoma, Kadoma-shi, Osaka-fu, Japan 571.

[10] Boreskov Institute for Catalysis in collaboration with the Budker Institute of Nuclear Physics in Novosibirsk.

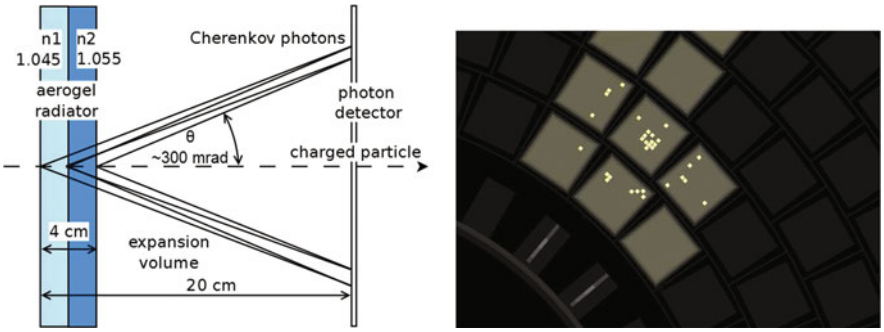


Fig. 7.13 (**a**) Proximity focusing RICH with two layers of the aerogel radiator: Cherenkov photons emitted in two aerogel tiles are detected on the same ring by the position sensitive photon detector, thus reducing the ring width. (**b**) Cosmic ray events registered by partially equipped detector. Reference [41]

The optical quality, light transmission T, see Fig. 7.12c, of aerogel is normally parameterized as:

$$T = T_0 \cdot \exp\left[-C \cdot \frac{t}{\lambda^4}\right] \tag{7.25}$$

where C is the clarity given in $\mu m^4/cm$ and t is the thickness in cm. T_0 describes the bulk properties of the aerogel and C the variation with the wavelength. The λ^4 term shows that the light attenuation, opacity κ,[11] is governed by Rayleigh scattering [40] which can be written as:

$$\kappa = \sigma_{\mathrm{Rlh}}(\lambda) \cdot N_0 \cdot t\,, \quad \text{where } \sigma_{\mathrm{Rlh}}(\lambda) \cong \frac{128\pi^5\zeta^2}{3\lambda^4} \cdot \frac{6+3\delta}{6-7\delta} \text{ for } \zeta = \frac{n-1}{2\pi N_0} \tag{7.26}$$

where N_0 is the number of particles per unit volume and δ is the polarization factor. δ is small and in the range from about 0.03 to 0.09.

The two-layer aerogel RICH detector of the Belle II spectrometer [41] will separate charged particles in the forward end-cap of the spectrometer inside a magnetic field of 1.5 T with a high separation capability in the momentum range from 0.5 to 3.5 GeV/c. See Fig. 7.13. The detector will be ready for data taking in 2018.

[11] Opacity is another term for the mass attenuation coefficient or, depending on context, mass absorption coefficient. κ_λ at a particular wavelength λ of the electromagnetic radiation.

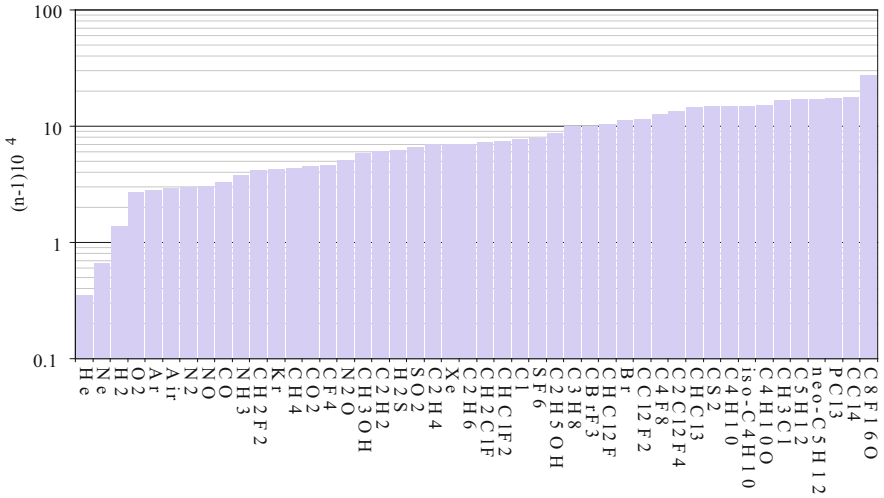


Fig. 7.14 Refractive index for some common fluids. D-line (589 nm). Data from [16, 22]

7.4.2.3 Fluids as Radiators

The relationship between the refractive index of a gas and the corresponding liquid, is given by:

$$\left[\frac{n^2-1}{n^2+2}\right]_{\text{gas}} = \left[\frac{p}{RT}\right]_{\text{gas}} \left[\frac{M}{\rho}\right]_{\text{liq}} \left[\frac{n^2-1}{n^2+2}\right]_{\text{liq}} \tag{7.27}$$

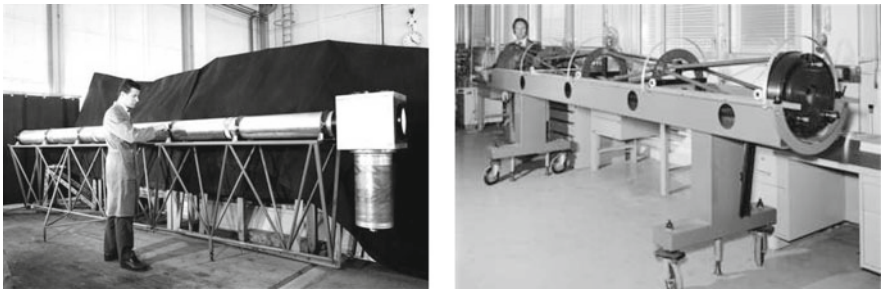
where p and T is the pressure and temperature of the gas, M and ρ is the molecular weight and density of the liquid and R is the gas constant (based on pressure and volume units $R = 82.0575$ (cm^3 atm)/(K mol)).

The refractive index for a number of fluids is plotted in Fig. 7.14.

7.4.3 Threshold Cherenkov Detectors

As soon as photon detectors, Chap. 3, coupled with the associated electronics, had the sensitivity to detect the low level of photons emitted through Cherenkov radiation, the first threshold Cherenkov detectors, see Figs. 7.15 and 7.16, were used in high energy experiments. The best known of these early experiments, is probably the discovery of the antiproton at the Radiation Laboratory of the University of California at Berkeley in 1955 [42].

The design of these threshold detectors is simple as is shown in Fig. 7.16a. In this sketch, the radiator is a gas. There is no problem to change it by inserting



(a) (b)

Fig. 7.15 (**a**) A threshold gas Cherenkov counter as used to tag particles in the secondary beams. CERN IT 6304088. (**b**) CEDAR counter (internal part). Here on the mounting bench. The counter is a differential Cherenkov, corrected for chromaticity, able to differentiate pions from kaons up to 350 GeV. Counters of this type were used in all SPS hadron beams. CERN PHOTO 7603033

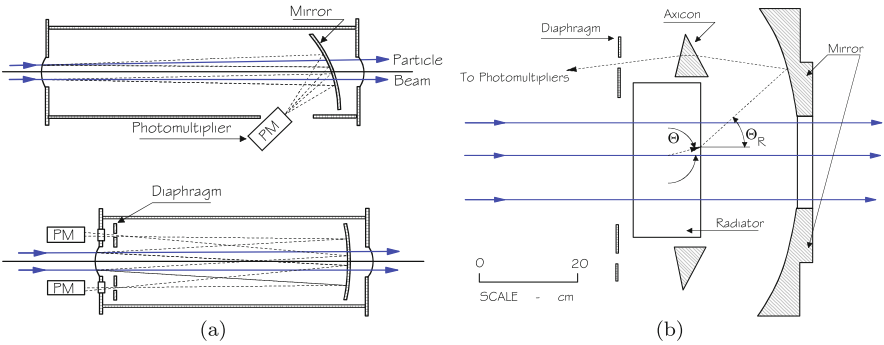


Fig. 7.16 (**a**) Top and bottom shows the working principles of respectively a threshold and a differential Cherenkov detector. (**b**) Is an achromatic liquid differential Cherenkov detectors, DISC: Differential Isochronous Self-Collimating; adapted from [43]

a solid or a liquid radiator, nor to change the pressure of the gas. It only affects the radiation length seen by the traversing particles. The solid angle covered by the detector is only limited by the design of the optics. A threshold Cherenkov detector can therefore be used both in the incoming beam to define the flavour of the primary particles as well as for identifying the secondaries. It should be noted that by introducing two, or more, detectors in series, positive particle identification can be achieved over a large momentum range.

A differential Cherenkov detector is shown in Fig. 7.16a. It is designed for a given value of the Cherenkov angle, such that:

$$\Theta = r/F \tag{7.28}$$

where r is the mean radius of the aperture of the diaphragm and F is the focal length of the mirror. The use of these detectors is mainly limited to parallel beams.

Assuming high energy particles and gas radiator, the resolution power can be written as:

$$\left[\frac{\Delta\beta}{\beta}\right]_{\text{limit}} = \tan\Theta \cdot \Delta\Theta \tag{7.29}$$

The coma[12] is the main error, given by:

$$\Delta\Theta_{\text{coma}} = \Theta^3 + \frac{\Theta^2}{4}\left[3\frac{b}{L} - \Theta\right] = \frac{3}{4}\Theta^3 \quad \text{if} \quad b \ll L \tag{7.30}$$

where b is the diameter of the incoming particle beam and L is the length of the gas radiator. The chromatic angular dispersion is given by:

$$\Delta\Theta_{\text{chrom}} = \frac{\Theta}{2\nu}\left[1 + \frac{1}{\gamma^2\Theta^2}\right] \quad \text{where} \quad \nu = \frac{n(\lambda_2) - 1}{n(\lambda_1) - n(\lambda_3)}, \tag{7.31}$$

representing the optical dispersion in the gas. λ_1 and λ_3 are the wavelengths appropriate for the limits of the spectral range. λ_2 is the mean wavelength. The total angular dispersion is then:

$$\Delta\Theta \approx \Theta^3 + \frac{\Theta}{2\nu}\left[1 + \frac{1}{\gamma_i^2\Theta^2}\right], \tag{7.32}$$

$i = 0, 1$ depending on the particle. We then get the limit for the maximum Cherenkov angle:

$$\Theta^4 + \frac{\Theta^2}{2\nu} \leq \frac{1}{2p^2}\left[m_1^2 - m_0^2 - \frac{m_i^2}{\nu}\right]. \tag{7.33}$$

For most applications, the Cherenkov angle will be smaller than this limit. The design will therefore be governed by the chromatic error.

To further diminish the errors, and thereby minimize $\Delta\beta/\beta$, a Differential Isochronous Self-Collimating, DISC, Cherenkov detector can be used. See Fig. 7.16b. With an optimized optics design a nearly achromatic condition can be achieved. That is,

$$\frac{\Delta\Theta(\lambda)}{\Delta\lambda} = 0 \quad \rightarrow \quad \frac{\Delta\beta}{\Delta\lambda} = 0 \tag{7.34}$$

[12] The aberration known as *coma* affects rays from points not on the axis of a lens. It is similar to spherical aberration in that both arise from the failure of the lens to image central rays and rays through outer zones of the lens at the same point. Coma differs from spherical aberration in that a point object is imaged not as a circle but as a comet-shaped figure (whence the term *coma*).

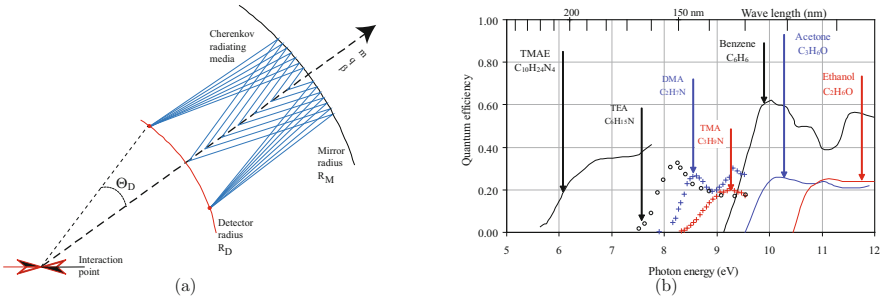


Fig. 7.17 (**a**) Ring imaging optics for particles emerging from a target or interaction region with zero impact parameter. The detected and emitted Cherenkov angles (Θ_D, Θ) are equal if the detector radius is correctly chosen.[45]. (**b**) The quantum efficiency for some photo sensitive vapours as function of photon energy. Adapted from [17]

Velocity resolution $\Delta\beta/\beta \sim 10^{-6} - 10^{-7}$ has been achieved [44]. These are very beautiful detectors, but with a somewhat limited usage as they require a near parallel beam, offer a limited solid angle and the material budget is not negligible.

7.4.4 Ring Imaging Cherenkov Detectors

The quest to make a ring imaging detector and thereby utilize all the inherent properties of Cherenkov radiation as described in Sect. 7.4, was long thwarted by the inability to get a high spatial resolution photon detector which was sensitive to single photons and compatible with photon absorptions, Fig. 7.8, in the media and photon transmission through windows, Fig. 7.9b. The breakthrough came in 1977 with the work of J. Séguinot and T. Ypsilantis [45, 46]. See Fig. 7.17a. Their work during the initial phase was mainly concentrated around MWPC, Chap. 4, and a photoionizing vapour additive to the chamber gas.

7.4.4.1 Photo Sensitive Vapours

Figure 7.17b shows the quantum efficiency for some photo sensitive vapours. The work with TEA,[13] Triethylamin $C_6H_{15}N$, and especially TMAE,[14] Tetrakis-(dimethylamino)-ethylene $C_{10}H_{24}N_4$ [47, 48], made it possible to work in the wavelength range from about 200 to 160 nm and thereby use fused silica as windows.

[13] http://webbook.nist.gov/cgi/cbook.cgi?ID=C121448&Units=SI.
[14] http://webbook.nist.gov/cgi/cbook.cgi?ID=C996703&Units=SI.

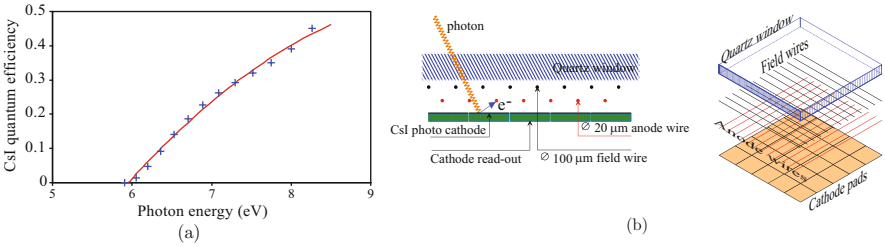


Fig. 7.18 (**a**) CsI quantum efficiency. (**b**) Sketch of a MWPC with CsI photo cathode

TMAE was the chosen photoionizing vapour, together with drift chambers, Chap. 3, for the first generation RICH detectors [49–51]. However, these fluids are difficult to handle and their usage is therefore now very limited. TEA and TMAE are chemically not reactive with respect to normal MWPC gases. They will, however, require an O_2 and water content of the carrier gas $\leq$ 10 ppm for stable operation. A drawback by using these molecules is the photon feedback. The photons created in the gas amplification process have a probability to convert. The main source of this background is from the ionization due to the charged particle going through the detector. The chambers were normally run at an amplification around $1 - 5 \cdot 10^5$ in order to be sensitive to single photons. The total probability for re-conversion thereby became larger than 1 and the chamber would break down. The number of feed-back photons can be written as $N_{\text{fp}} = \iota \cdot G$ where G is the total chamber gain.[15] $\iota \sim 7 - 8 \cdot 10^{-6}$ in CH_4 due to photon absorption for wavelengths below 143 nm. See Eq. (7.21).

A number of ingenious chamber designs were made to minimize the photon feed-back. The designs are a compromise between detection efficiency, ease of operation and fabrication and drift of electrons in a $\mathbf{B}\times\mathbf{E}$ configuration. Even at stable operating conditions, some photons will escape and give rise to an event correlated background. This background is difficult to disentangle from the real signal in high occupancy events and particularly with TMAE due to its long photon conversion length.

7.4.4.2 CsI Photo Cathode

The next step in high spatial granularity, or pixilated, photon detectors for RICH came with the CsI photon detector [52]. CsI is an alkali halide crystal which has a good quantum efficiency, Fig. 7.18a, below 200 nm and is stable in normal dry and O_2 free chamber gases [53]. The development was triggered by the need for a faster detector at the arrival of LHC and similar accelerators. In a MWPC structure,

[15] The measured chamber gain might be smaller due to charge sharing and electronics time constants.

the CsI photo cathode can either be deposited as a reflective, Fig. 7.18b, or as a semi-transparent layer [54]. The latter would, in the case of Fig. 7.18b lay-out, be a layer on the quartz window. The maximum quantum efficiency for semi-transparent CsI is for a thickness of about 11 nm in the wavelength range from 210 to 170 nm. The thickness does not matter for a reflective photo cathode and is normally in the range of 150–200 nm. A semi-transparent CsI photo cathode will have a quantum efficiency of about 0.7 compared to a reflective one. It should be noted that the photon conversion efficiency is strongly depending on the bulk structure and morphology of the CsI layer; that is, the roughness of the substrate and the connectivity of the layer. Particularly thin layers can become a collection of unconnected islands. Post-production heat treatment has proven advantageous.

As for the photosensitive vapours, a CsI photo cathode will be sensitive to the photon feed-back from the gas amplification process, see Sect. 7.4.4.1. A stable operation of the chamber is therefore a compromise between single photon efficiency, electronics sensitivity, signal shaping and gas amplification.

As few, if any other photon detector, can beat a gas based detector in cost efficiency and geometrical acceptance, a number of similar, but not identical, detector set-ups are proposed and investigated. The main emphasis is on limitation of photon feed-back, on better and more stable photo cathodes and on time resolution. This work is also partially driven by very large Cherenkov detectors for astrophysics. A very promising research and development is in gaseous micro pattern detectors with Bialkali photo cathodes. We will not discuss these here, but refer the reader to [55]. An overview of the current status and perspectives of gaseous photon detectors can be found in [56].

7.4.4.3 Vacuum Based Photon Detectors

The working principles of vacuum based photon detectors like photo multiplier tubes are discussed in Chap. 3. Although small diameter PM tubes, diameter 10 mm upwards, have been used for a long time in Cherenkov detectors, cost, balanced with space resolution and material budget, made them less attractive. The introduction of multi anode and pixilated silicon anode detectors, together with fast and sensitive electronics changed this. The first generation of multi anode photo tubes required a lens system [57] in order to give good geometrical acceptance. See Fig. 7.19a.

The schematic of a Hybrid Photon Detector [60, 61], HPD, is shown in Fig. 7.19b. In these detectors the encapsulated pixilated silicon detector is bump-bonded onto the read-out electronics. The capacitance is thereby small and the associated noise low. It also requires only a few vacuum feed-throughs. The photo cathode is normally a S20.[16] Under the influence of the electric field, the photo-electron is accelerated onto the silicon detector. In the example given in Fig. 7.19b, the 20 kV potential between anode and cathode gives a cross-focusing field with a

[16] S20 is a tri-alkaline (Sb-Na-K-Cs) semi-transparent photo cathode.

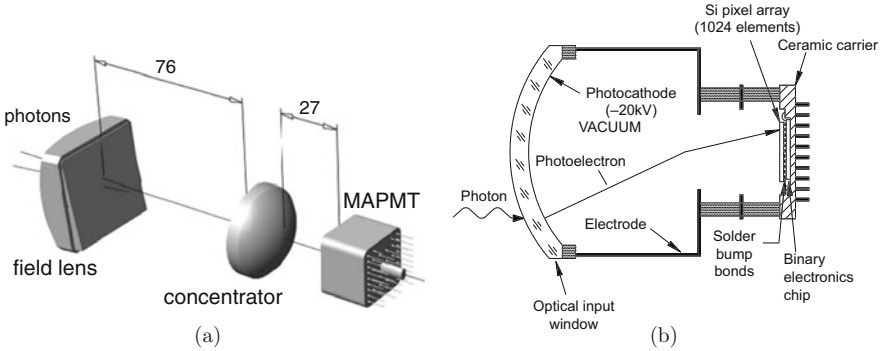


Fig. 7.19 (**a**) Optical arrangement of the COMPASS MAPMT and the fused silica lens telescope. With permission [58]. (**b**) Schematic arrangement of the LHCb Hybrid Photon Detector. With permission [59]

demagnification of ∼5. Other field configurations can be used [61]. The granularity of the silicon detector can be tailored as function of the required geometrical resolution.

These new photon detectors with a maximum quantum efficiency of about 30–35% around 300 nm, have made the choice of Cherenkov radiators and photon windows much more flexible. It has for instance allowed the use of aerogel in Ring Imaging Cherenkov detectors. See Sect. 7.4.2.2 and Fig. 7.12c.

Current research and development is mainly concentrated on faster and cheaper detectors with large geometrical acceptance. These are detectors like silicon avalanche photo diodes, micro channel plates and large area flat panel multi-anode PMTs. The reader is referred to Chap. 3.

An overview of the current status and perspectives of vacuum-based photon detectors can be found in [62].

7.4.5 Optics

We can broadly divide the light collection system of Ring Imaging Cherenkov detectors into two distinctive classes.

- Proximity focusing, or direct light collection as in Fig. 7.20.
- Concave mirrors as in Fig. 7.17a in Sect. 7.4.4.

7.4.5.1 Proximity Focusing

In the first case with proximity focusing optics, the resolution relies on the thinness, l, of the radiator in comparison to the expansion length, L. That is, $l \ll L$. The Cherenkov light will then describe a thin cone around the charged particle and

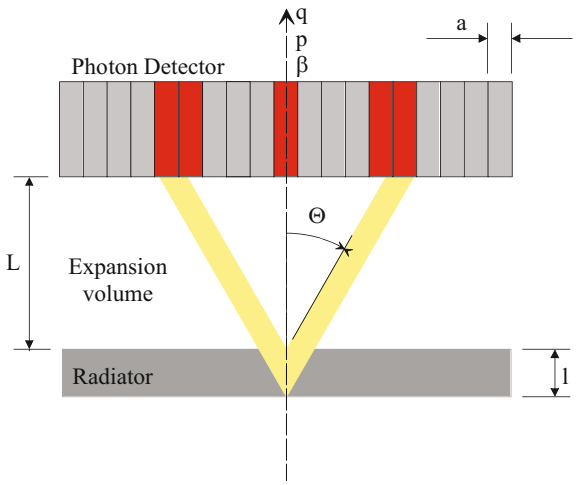


Fig. 7.20 Proximity focusing arrangement

will give rise to a finite width, conic section image where the detector plane intercepts this cone. If the particle is not perpendicular to the radiator and the detector planes, this circular image becomes distorted to an elliptic or a hyperbolic image. Depending on the refractive index of the radiator, photon window material and the expansion gap, light might be trapped due to total internal reflections. See Sect. 7.4.2.1. As the photon detector has to be placed in the path of the charged particle, the material budget may become prohibitive. However, this detector configuration is well adapted to 4π detectors with high refractive index radiators [51, 63].

7.4.5.2 Focusing Mirrors

Detectors which cover large solid angles require large focusing mirrors as in Fig. 7.21. There are two.[17] options, parabolic [65] and spherical [66, 67] mirror. The choice of mirror substrate is a balance between cost, ease of fabrication and performance. Whereas the material budget is normally not an issue in astrophysics, see Sect. 6.1 and Fig. 7.21a and b, it is one of the main concerns in accelerator based experiments as the mirrors must be inside the acceptance. If spherical aberration becomes a dominant contribution to the total error in the Cherenkov angle calculation, parabolic mirrors should be used.

[17]We will not discuss here ellipsoid nor hyperbolic mirrors. For correctors like Schmidt and Maksutov, the reader is referred to [40].

(a) (b) (c)

Fig. 7.21 (**a**) and (**b**) The VERITAS Telescope 1 as installed at the Whipple Observatory base camp. The collector dish has a diameter of 12 m and a focal length of 12 m and comprises 350 mirror facets. A 499-PMT camera is installed in the box at the focal point. Courtesy of the VERITAS Collaboration [64]. (**c**) COMPASS [67] mirror wall of RICH 1. CERN EX 0106007 01

Table 7.3 Basic material properties for some mirror substrates together with substrate rigidity, K, and the rigidity divided by material thickness in units of radiation length

Material	X_0 [cm]	E [10^4 MPa]	α [10^{-6}/°C]	Relative rigidity K	K/X_0 relative
Beryllium	35.3	28.9	11.3	1	1
Plexiglas	34.4	0.33	70	0.012	0.011
Pyrex glass	12.7	6.17	3.2	0.213	0.076
Aluminium	8.9	6.9	23.9	0.238	0.060

X_0 is the radiation length, E the Young's module and α is the coefficient of thermal expansion

The material option for the mirror substrate is a balance between radiation length, size of the substrate and stability. Some options[18,19] are given in Table 7.3.

The rigidity, K, of a thin mirror substrate is roughly given by:

$$K \propto \frac{Et^3}{D^2} \tag{7.35}$$

where E is Young's modulus, t is the substrate thickness and D is the diameter. The superiority of substrate materials like beryllium is clear in Table 7.3. In this table

[18]Plexiglas is Poly(methyl methacrylate) (PMMA) by Evoniks Business Unit Performance Polymers.

[19]Pyrex, Corning Incorporated, is made of 4% boron, 54% oxygen, 3% sodium, 1% aluminium, 38% silicon, and less than 1% potassium.

substrates of diameter 500 and 5 mm thickness are compared. However, beryllium is not a good reflector nor a good support for a reflecting surface. A thin glass face is therefore required on the beryllium as support for the reflector [68]. This glass surface can also be used to adjust the focal length of the mirror. The main challenge is to use a glass which has the same thermal expansion coefficient as beryllium.

Thin and robust mirror substrates can be made as a sandwich assembly. The kernel is normally a honeycomb or foam and the inner and outer skin are preformed to about the right radius of curvature. The final adjustment is done at the assembly stage or by reshaping, by polishing, the reflecting skin later. The skin can be high strength carbon fibre sheets [69], easily formed Plexiglas [70] or simple metal structures [71]. Glass with glass-foam kernel has also been built [72].

Glass is still the most used substrate for mirrors. It is easily shaped and machined and the ageing behaviour is well known. Stresses in the material can be simply relieved. It is also inert in most Cherenkov radiators. It is normally slumped to the required shape and then polished to the final focal length. Its principal drawback is the radiation length.

7.4.6 *The Reflective Surface*

The reflectivity of a surface is a function of the incident angle and energy of the light and the dielectric structure of the surface. The principle is discussed in [40] and more specifically in [73]. See Fig. 7.22a. A high reflectivity layer is over-coated by one or more transparent films of high and low refractive indices. Aluminium and silver are good reflectors with peak reflectivity of respectively ∼92% and ∼96%. Aluminium, the most widely used metal for reflecting films, offers consistently high reflectance throughout the visible, near-infrared, and near-ultraviolet regions of the spectrum.

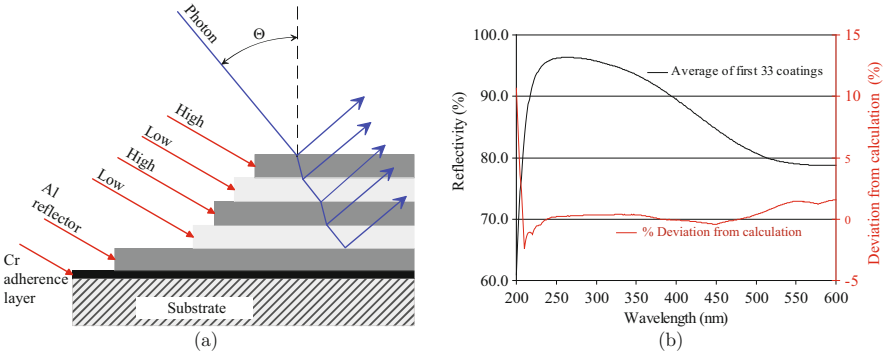


Fig. 7.22 (**a**) Schematic representation of a metal multi-dielectric mirror [73]. (**b**) Measured and calculated reflectivity of a multi-dielectric mirror coating. The stack is Cr-Al-SiO_2-HfO_2. Adapted from [73]

Table 7.4 Typical process parameters for a multi-dielectric mirror coating [73]

Material	Purity [%]	Chamber pressure [Pa]	Deposition rate [nm/s]	Thickness (geom.) [nm]
Cr	99.98	$2 \cdot 10^{-5}$	1	20
Al	99.999	$2 \cdot 10^{-5}$	5	85
SiO_2	99.99	10^{-3} O_2	0.2	28
HfO_2	99.9	10^{-3} O_2	0.2	38

While silver exhibits slightly higher reflectance than aluminium through most of the visible spectrum, the advantage is temporary because of oxidation tarnishing. Aluminium also oxidizes, though more slowly, and its oxide is tough and corrosion resistant. Oxidation significantly reduces aluminium reflectance in the ultraviolet and causes slight scattering throughout the spectrum. Generally, all reflective layers need a protective film.

Material like SiO_2 and MgF_2 have low refractive index in comparison to HfO_2 and TiO_2. Properties like residual stress, adherence, resistivity to abrasion and humidity and coating yield are essential in the selection process for these layers. The optical thickness of the layers, $d_{\text{opt}} \propto \cos\Theta$, is normally chosen to be $\lambda/4$. A dielectric coating will lead to a wavelength and angle dependent modulation of the reflectivity. The larger the ratio between the refractive indices in a Low/High pair, the higher is the peak reflectivity and width of the enhanced region. Adding more pairs for the same wavelength range, will enhance the peak reflectivity, but narrow the wavelength range. The layer stack will normally be terminated with a high refractive index layer. In this way the mirror reflectivity can be optimized for the wavelength range of the photon detector.

Mathematically approximation codes[20] will predict the behaviour of the multilayer film. The accuracy only depends on the knowledge of the refractive index and the absorption in the deposited layers. These optical properties are however dependent of the deposition method and processing parameters.

An example is shown in Fig. 7.22b. Layers of Cr, Al, SiO_2 and HfO_2 are used on a glass substrate. See Table 7.4 for process parameters. This coating is optimized for a wavelength of 275 nm in order to match a S20, footnote 16, photo cathode and compared with calculations. See footnote 20 for the calculation.

7.4.6.1 Mirror Imaging Quality

The error introduced by the imaging quality of a RICH mirror should be small compared to all other errors in the detector. If the mirror is a perfect spherical

[20] FilmStar Design, FTG Software Associates, Princeton, NJ,
SCI Film Wizard, Scientific Computing International, Carlsbad, CA
or similar.

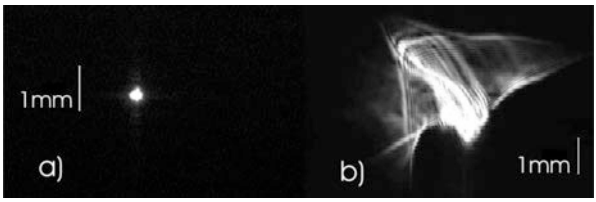


Fig. 7.23 (**a**) Spot image for a high precision glass mirror. (**b**) Spot image for a thin glass mirror [66]

surface, the spot on the focal plane would have the size given by the diffraction limit. For a circular mirror of diameter D and a radius of curvature R, the diffraction limited spot diameter, d, at the third maximum, corresponding to 95.3% of the focused light, is given by:

$$d = 2R\tan\alpha \quad \text{for} \quad \sin\alpha = \frac{\lambda x}{\pi D} \quad \text{and} \quad x = 3.7\pi \tag{7.36}$$

For a wavelength $\lambda = 641\,\text{nm}$,[21] $D = 0.50\,\text{m}$ and $R = 8\,\text{m}$, $d = 76\,\mu\text{m}$.

Real mirrors have real imperfections. Fig. 7.23 shows the difference between a high precision and a thin glass mirror. The mirror in Fig. 7.23a is a 50 mm thick glass mirror of diameter 400 mm and a radius of curvature of 7.8 m. The Fig. 7.23b mirror is 7.5 mm thick with a diameter of 400 mm and a radius of curvature of 7.8 m. 95% of the focused light for the first mirror is inside circle of diameter 0.23 mm. The corresponding diameter for the second mirror is 3.4 mm. This mirror also features irregularities at the edges of the surface. The average quality of a mirror is well described by the spot size at the focal plane and is normally sufficient as a qualification parameter. Let D_0 be the diameter of this spot which encompasses 95% of the light. $\sigma_s = D_0/4$ is the RMS of the light distribution if this distribution was Gaussian. The error induced by the mirror is then given by:

$$\sigma_\Theta = \frac{\sqrt{\sigma_s^2 + \sigma_p^2}}{2R} \approx \frac{\sigma_s}{2R} = \frac{D_0}{8R} \tag{7.37}$$

where σ_p is the resolution of the point source.

The determination of the spot shape can be an invaluable tool in the development and fabrication process. The quantification of the variation in the radius of curvature across a substrate can be used to improve the resolution of the system. It can be particularly important for large mirrors.

Shack-Hartmann sensors, Ronchi test method, Foucault method and similar measurement methods are described in detail in [74]. We will only show the power of these methods with one example.

A sketch of a Ronchi test set-up is shown in Fig. 7.24a. A beam of coherent, quasi-monochromatic light is brought to focus by an optical system that is under-

[21] Red laser diode.

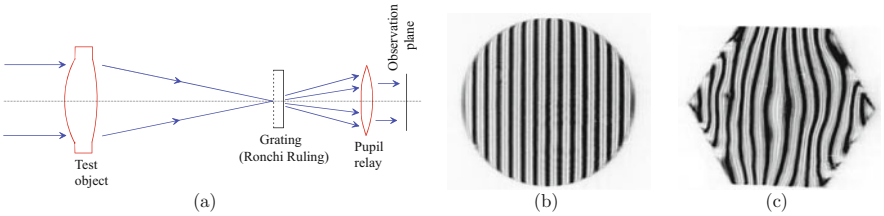
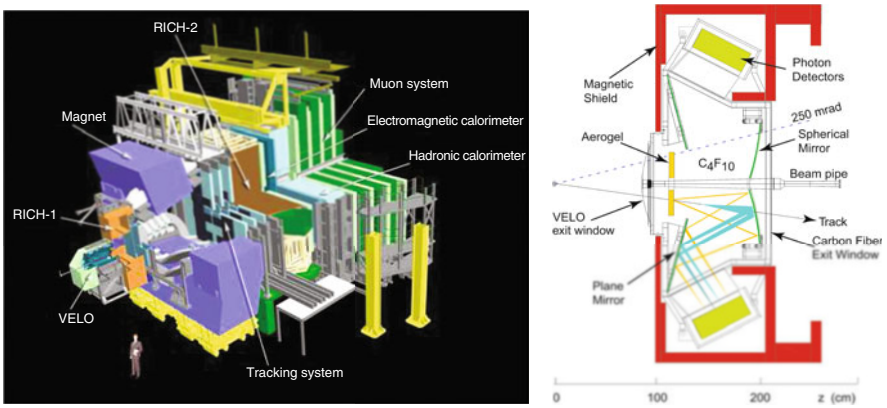


Fig. 7.24 (**a**) General set-up for a Ronchi test. (**b**) Ronchigram of a high precision spherical glass mirror. Thickness 50 mm. (**c**) Ronchigram of a thin spherical glass mirror. Thickness 4.5 mm. Ronchi ruling 1 mm

going tests to determine its aberrations. A lens, or more generally any optical system consisting of an arrangement of lenses and mirrors, is placed in the position *Test Object*. A diffraction grating, placed perpendicular to the optical axis in the vicinity of the focus, breaks up the incident beam into several diffraction orders. The diffracted orders propagate, independently of each other, and are collected by a pupil relay lens, which forms an image of the exit pupil of the object under test at the observation plane. For a concave mirror, deviation from a spherical surface will result in deformation of the fringes. The measurement is only sensitive to changes in radius of curvature perpendicular to the grating direction. Results are shown in Fig. 7.24. Figure 7.24b is a Ronchigram for a high precision spherical mirror, whereas Fig. 7.24c is for a thin large mirror. For the first mirror, the interference lines are straight which shows that the deviation from the ideal shape is smaller than the resolution of the Ronchi ruling. For the second mirror, the interference lines are distorted. In the centre, the lines bow outward and indicate parabolic deformation. On the edges, the lines bow inward to indicate an oblate spheroid surface.

7.4.7 Ring Finding and Particle Identification

As explained in Sect. 7.4, Cherenkov light is produced in a cone at polar angle Θ_C relative to the particle trajectory, as given by Eq. (7.8) for a particle travelling at velocity β. In a RICH detector the light is focussed onto a detector plane as a ring image. For the classical RICH geometry illustrated in Fig. 7.17a and [46], the detected photons corresponding to a track passing through the detector would form a circular ring image centred on the track impact point on the detector. The issues discussed in this section are the finding of the ring, i.e. the pattern recognition to associate the detected photons to a given track, and the particle identification, i.e. the determination of the particle type, given the photons that are associated to its track. Examples are taken from LHCb, Fig. 7.25, the dedicated B physics experiment at the LHC, which has two RICH detectors [75]. A review of other approaches can be found in [76].



(a) (b)

Fig. 7.25 (**a**) View of the LHCb detector. (**b**) Side view schematic layout of the RICH 1 detector. Reference [75]

For the simple detector geometry of Fig. 7.17a, and for a single track passing through the detector, the circular image implies that the photons from the track all lie at a constant radius on the detector plane, when measured from the track impact point. The radius r is related to the Cherenkov angle, Θ_C, by:

$$r = R\Theta_C/2 \tag{7.38}$$

where R is the radius of curvature of the spherical focussing mirror. For a given track the pattern recognition could therefore simply be performed by plotting the radius of all photons in this way, and searching for a peak in the distribution. Due to the finite resolution, this signal peak will have a roughly Gaussian shape, with width corresponding to the resolution. Sources of finite resolution include the pixel size of the photon detector, and the fact that the refractive index has some dependence on the photon wavelength, leading to a chromatic term in the resolution. Background hits that are distributed randomly across the detector plane, for example from noise in the photon detector, will appear as a contribution in the plot of detected photon radius that increases roughly linearly with radius (due to the increasing area swept out on the detector plane as the radius increases). This situation is illustrated in Fig. 7.26a.

Given the reconstructed radius r, the Cherenkov angle can be calculated from Eq. (7.38), and thus the velocity β of the particle determined from Eq. (7.8). To make the final step of identifying the particle, the momentum p must also be known, usually from the tracking system of the experiment that measures the curvature of the track in a magnetic field. Then the mass m of the particle can be determined using relativistic kinematics:

$$m^2 = p^2(\beta^{-2} - 1)/c^2 \tag{7.39}$$

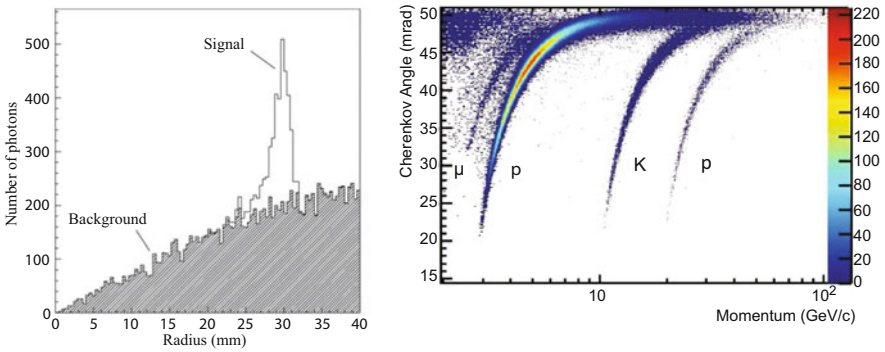


Fig. 7.26 (**a**) Distribution of photons in radius around the track, for a set of tracks in one of the LHCb RICH detectors; the peak from the photons associated to the track is visible, along with background from other sources. (**b**) Reconstructed Cherenkov angle for isolated tracks, as a function of track momentum in the C_4F_{10} radiator. The Cherenkov bands for muons, pions, kaons and protons are clearly visible. Reference [77]

and this identifies the particle type. An example is shown in Fig. 7.26b where the reconstructed Cherenkov angle has been plotted versus momentum for all the particles in a set of events, and the loci of points corresponding to particles with different masses are clearly seen.

In practical implementations of the RICH technique, the optical system usually differs from the simple classical layout, so as to avoid the material of the photon detectors being placed within the acceptance of the spectrometer. For example, the RICH detectors of the LHCb experiment involve a spherical focussing mirror that is tilted with respect to the track direction, and an additional planar mirror to bring the Cherenkov light to photon detectors sited outside the acceptance, while limiting the overall size of the detector system. This complicates the reconstruction somewhat, as the ring images are no longer circular but become distorted into roughly elliptical shapes, and the track no longer passes through the detector plane, but its image on that plane has to be calculated from knowledge of the optics. There is also an additional contribution to the resolution, due to the spherical aberration resulting from imaging the photons from off-axis tracks, but this can usually be arranged to be smaller than the limiting chromatic effect. The distortion of the ring image can be exactly corrected for by reconstructing the Cherenkov angle for each photon-track pair. For a spherical focussing mirror the analytical solution of this calculation involves the solution of a quartic equation. See [78]. For reasons of speed, a numerical approach can be used instead, ray-tracing photon candidates through the optical system and calibrating the distortion of the ring image in this fashion. The peak search is then performed in Cherenkov angle space, rather than radius on the detector plane.

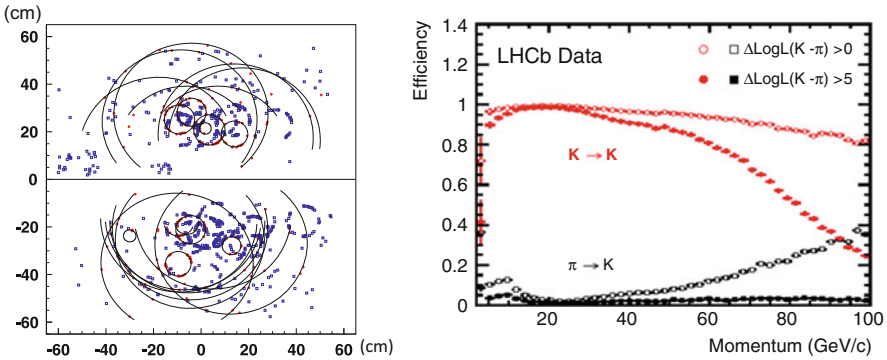


Fig. 7.27 (**a**) Ring images from tracks passing through the RICH 1 detector of LHCb, from a single proton-proton collision event at the LHC. (**b**) Kaon identification efficiency and pion misidentification rate as measured using data. Two different $\Delta\log\mathcal{L}$ (K-π) requirements have been imposed on the samples, resulting in the open and filled marker distributions, respectively. Reference [77]

This approach of peak searching works well in situations of low track multiplicity, where the ring images from tracks are well separated. However, at the LHC the track density is high, as illustrated for a typical event in Fig. 7.27a. In this case the main background to the reconstruction of the ring image of a given track comes from the overlapping rings from other tracks. It is therefore advantageous to consider the optimization of photon assignment to all of the tracks in the event simultaneously, in a so-called global approach. Since a momentum measurement is required to convert a measured ring image into particle identification, as discussed above, it makes sense to use the reconstructed tracks in the event as the starting point for pattern recognition. Trackless ring searches have been developed, but are mostly relevant for background suppression, rather than particle identification [76]. Furthermore, the number of stable charged particle types that are required to be identified is rather limited, typically five: e, μ, π, K, p. The pattern recognition can be made faster by just searching for these particle types, i.e. hypothesis testing. For applications where speed is crucial, such as use in the trigger of the experiment, the number of hypotheses compared can sometimes be further reduced, depending on the physics process that is being selected, e.g. simply comparing π and K hypotheses [79]. On the other hand, if one is interested in an unbiased search for charged particles (such as exotic states) then alternative approaches exist that do not rely on preselected hypotheses [80].

The pattern recognition then proceeds by taking the most likely hypothesis for each of the tracks in the event, typically the π hypothesis as they are the most abundantly produced particle (at the LHC). The likelihood is then calculated that the observed pattern of photons was produced by the particles, under these first choices of mass hypotheses. Conceptually this corresponds to taking the product of terms for each photon according to how close it is to the nearest ring image, assuming

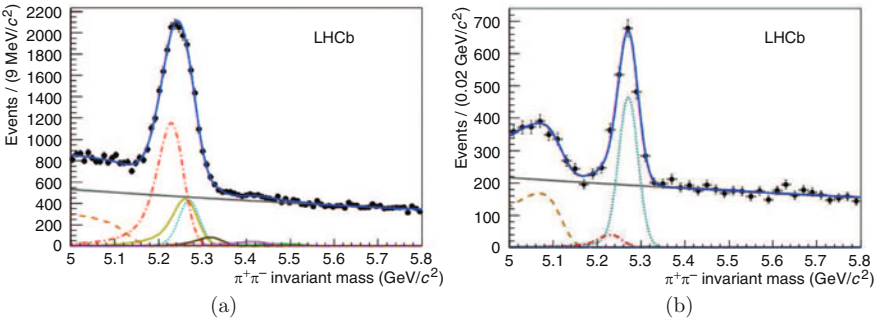


Fig. 7.28 (**a**) Invariant mass distribution for B→ h^+h^- decays in the LHCb data before the use of the RICH information, and (**b**) after applying RICH particle identification. The signal under study is the decay $B^0 \to \pi^+\pi^-$, represented by the turquoise dotted line. The contributions from different b-hadron decay modes ($B^0 \to K\pi$ red dashed-dotted line, $B^0 \to$ 3-body orange dashed-dashed line, $B_s \to$KK yellow line, $B_s \to K\pi$ brown line, $\Lambda_b \to$pK purple line, $\Lambda_b \to p\pi$ green line), are eliminated by positive identification of pions, kaons and protons and only the signal and two background contributions remain visible in the plot on the right. The gray solid line is the combinatorial background. Reference [81]

a Gaussian probability distribution around each ring. A term is also added to the likelihood from the comparison of the total number of photons assigned to a track, compared to the expected number given the mass hypothesis and momentum. The tracks in the event are then all checked to see which would give the greatest increase in the total likelihood of the event, if its hypothesis were to be changed, and the mass hypothesis of the one giving the greatest increase is then changed. This procedure is iterated until no further improvement in the likelihood can be achieved, at which point the maximum-likelihood solution to the pattern recognition has been found. By the use of various computational tricks [78] this algorithm can be reasonably fast, typically taking a similar CPU time to the track finding algorithm. The performance of this approach to particle identification when applied to LHCb events (of the type shown in Fig. 7.27a) is illustrated in Fig. 7.27b. The efficiency for identifying kaons and the misidentification rate of pions are both shown as a function of momentum.

An example from the LHCb experiment of the resulting powerful particle identification in B→ h^+h^- decays is shown in Fig. 7.28. The LHCb experiment moves to a fully software trigger where the RICH information is embedded.

7.5 Transition Radiation Detectors

A charged particle in uniform motion in free space will not radiate. It can radiate if it traverses a medium where the phase velocity of light is smaller than the velocity of the charged particle. This is Cherenkov radiation as discussed in Sect. 7.4 and was first correctly described by P.A. Cherenkov and S.I. Vavilov in 1934 and formulated

by I.M Frank and I.E. Tamm in 1937 [15]. This radiation was worked into the Bethe-Bloch formalism in 1940 by E. Fermi, see Chap. 2 and [82].

There is another type of radiation when the charged particle traverses a medium where the dielectric constant, ε, varies. This is transition radiation. It is analogous to bremsstrahlung. In both cases the radiation is related to the phase velocity of the electromagnetic waves in the medium and the velocity of the particle. In the case of transition radiation, the phase velocity changes whereas the particle velocity changes for bremsstrahlung. Transition radiation is, like bremsstrahlung, strongly forward peaked.

V.L. Ginzburg and I.M. Frank predicted in 1944 [83] the existence of transition radiation. Although recognised as a milestone in the understanding of quantum mechanics, transition radiation was more of theoretical interest before it became an integral part of particle detection and particle identification [84].

The exact calculation of transition radiation is complex and we will not repeat the mathematics here. The reader is referred to [18, 85, 86]. Specific discussions can be found in [87, 88]. We will here just recall some of the central features.

Transition radiation is emitted when a charged particle traverses a medium with discontinuous dielectric constant. Let $[\mathbf{E}_1, \mathbf{H}_1]$ be the Lorentz transformed Coulomb field of the charged particle in medium 1 and $[\mathbf{E}_2, \mathbf{H}_2]$ the corresponding one in medium 2. See Fig. 7.29a. $[\mathbf{E}_1, \mathbf{H}_1]$ and $[\mathbf{E}_2, \mathbf{H}_2]$ do not match at the boundary. In order to satisfy the continuity equation, a solution of the homogeneous Maxwell equation must be added in each medium. This is the transition radiation. The angular distribution of transition radiation by a perfectly reflecting metallic surface is of the form:

$$J(\Theta) = \omega \frac{dN}{d\omega d\Omega} = \frac{\alpha}{\pi^2} \left(\frac{\Theta}{\gamma^{-2} + \Theta^2} \right)^2 \tag{7.40}$$

where $\gamma = E/m \gg 1$ in natural units, $\hbar = c = 1$, $\alpha \simeq 1/137$ is the fine structure constant and $\Theta \ll 1$ is the angle of the photon with respect to the velocity vector $\mathbf{v}$ of the charged particle. Θ is along $\mathbf{v}$ for forward transition radiation or its mirror

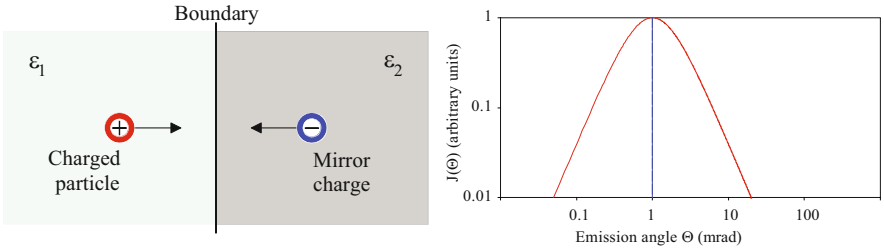


Fig. 7.29 (**a**) Schematic representation of the production of transition radiation at a boundary. (**b**) Transition radiation as function of the emission angle for $\gamma = 10^3$. Eq. (7.40)

direction for backward transition radiation. N is the total number of emitted photons. Equation (7.40) is plotted in Fig. 7.29b.

The energy radiated from a single surface, assuming $\varepsilon_0 \to \varepsilon$, is given by:

$$W = \frac{1}{3}\alpha Z^2 \omega_p \gamma \tag{7.41}$$

where ω_p is the plasma frequency.

7.5.1 Plasma Frequency

The influence of the plasma frequency was shown in the saturation of the relativistic rise expressed by the Bethe-Bloch formula, Chap. 2 and [82], due to the polarisation of the medium:

$$\frac{\delta}{2} = \ln\frac{\omega_p}{I} + \ln\beta\gamma - \frac{1}{2} \tag{7.42}$$

where I and ω_p are respectively the mean excitation energy and the plasma frequency of the medium and δ is the density correction.

The plasma frequency, ω_p, is the natural frequency of density oscillations of free electrons and its value depends only weakly on the wavelength. Longitudinal plasma waves are resonant at ω_p. Transverse electromagnetic waves are absorbed below ω_p. If $\omega < \omega_p$, the index of refraction has an imaginary part and the electromagnetic waves are attenuated or reflected. If $\omega \gg \omega_p$, the index is real and a metal becomes transparent. For large ω one can write

$$n^2 = 1 - \left(\frac{\omega_p}{\omega}\right)^2 \tag{7.43}$$

The plasma frequency is given as:

$$\omega_p^2 = \frac{NZe^2}{\varepsilon_0 m} \tag{7.44}$$

and depends only on the total number, NZ, of free electrons per unit volume. The plasma frequency can be approximated with:

$$\omega_p(\text{eV}) \simeq 28.8\sqrt{\frac{\rho(\text{g/cm}^3)\cdot z}{A}} \tag{7.45}$$

where z is the effective number of free electrons per unit volume. Table 7.5 gives the corresponding calculated and measured wavelength, λ_p, for alkali metals. $z = 1$ for alkali, group 1a, metals. The calculated plasma energies in Si, Ge and InSb are

Table 7.5 Ultraviolet transmission limits of alkali metals in nm [89]

Material	A	Z	λ_p [nm]	
			Calculated	Measured
Li	6.939	3	155	155
Na	22.99	11	209	210
K	39.10	19	287	315
Rb	85.47	37	322	340
Cs	132.95	55	362	–

Table 7.6 Radiator material properties [90]

Material		ρ [g/cm^3]	ω_p [eV]	X_0 [cm]
Lithium		0.534	13.8	148
Beryllium		1.84	26.1	34.7
Aluminium		2.70	32.8	8.91
Polyethylene	CH_2=CH_2	0.925	20.9	49
Mylar	$C_5H_4O_2$	1.38	24.4	28.7
Air		$2.2 \cdot 10^{-3}$	0.7	$30.9 \cdot 10^3$

based on four valence electrons per atom. In a dielectric the plasma oscillation is physically the same as in a metal: the entire valence electron sea oscillates back and fourth with respect to the ion core. Table 7.6 tabulates properties of some commonly used radiator material.

7.5.2 *Formation Zone*

A minimum thickness is required in order to efficiently produce the transition radiation as the evanescent field has a certain extension. This is the formation zone and is illustrated in Fig. 7.30 for a stack of aluminium, $\omega_p(\mathrm{Al}) \sim 32.8\,\mathrm{eV}$, and air, $\omega_p(\mathrm{air}) \sim 0.7\,\mathrm{eV}$. The length of the formation zone, d, can be written as:

$$d = \frac{2c}{\omega}\left[\gamma^{-2} + \Theta^2 + \left(\frac{\omega_p}{\omega}\right)^2\right]^{-1} \tag{7.46}$$

which has a maximum, $d_{\max}$, at $\omega = \gamma\omega_p/\sqrt{2}$ for $\Theta = \gamma^{-1}$, which is equivalent to the maximum intensity as can be seen from Eq. (7.40) and Fig. 7.29b.

$$d_{\max}(\mu\mathrm{m}) \sim 140 \cdot 10^{-3}\frac{\gamma}{\omega_p(\mathrm{eV})} \tag{7.47}$$

Inserting Eq. (7.45) in Eq. (7.47), we see that for media with a density in the order of 1, $\omega_p \simeq 20\,\mathrm{eV}$ and $\mathrm{d_{max}} \simeq 7\,\mu\mathrm{m}$ for $\gamma = 1000$. For a gas, ω_p is about 30 times smaller due to the reduced density and $\mathrm{d_{max}}$ thereby 30 times longer for same γ.

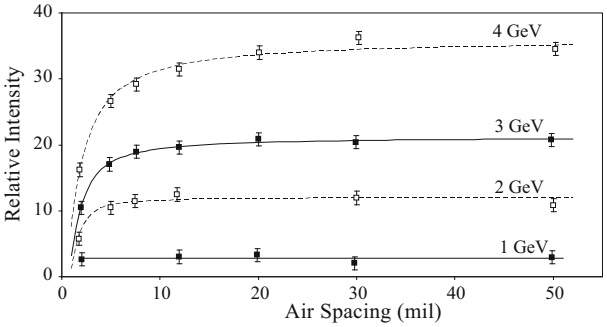


Fig. 7.30 Relative intensity of transition radiation for different air spacing. Each radiator is made of 231 aluminium foils 1 mil thick. (1 mil = 25.4 μm). Particles used are positrons of 1–4 GeV energy (γ = 2000–8000). Adapted from [91]

Using numbers for the experimental set-up in Fig. 7.30, we get $d_{max} \sim 1.5$ mm for $\gamma = 8000$.

7.5.3 Transition Radiation Detectors

From the discussion above, transition radiation can be characterized by the following:

- Transition radiation is a prompt signal.
- Transition radiation is not a threshold phenomenon.
- The total radiated power from a single interface is proportional to γ.
- The mean emission angle is inversely proportional to γ.

In general terms, there are two different types of transition radiation detectors:

1. The detectors working in the low energy, optical, range.
2. The detectors working in the X-ray range.

We will briefly introduce the first one and use a little more space on the second class of X-ray transition radiation detectors.

7.5.3.1 Optical Transition Radiation Detectors

J.E. Lilienfeld [92] was probably the first,[22] in 1919 to observe that in addition to X-rays, radiation ranging from visible light through the ultraviolet is emitted

[22] This statement has been contested over the years and could be due to a confusion between transition, Cherenkov radiation and bremsstrahlung. See [93].

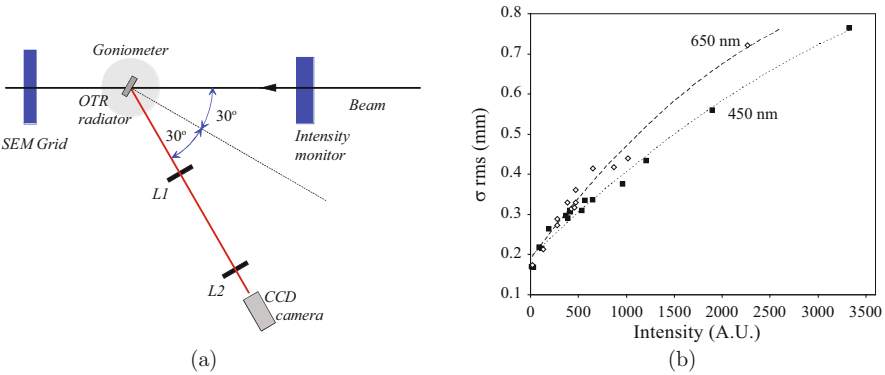


Fig. 7.31 (**a**) Sketch of an experimental set-up for measurement of optical transition radiation with secondary emission, SEM, grid and beam intensity monitor. The transition radiation foil is tilted by 30° with respect to the beam line. The optical system is defined by two lenses and a CCD camera. (**b**) Measured rms beam size values as a function of the total intensity for $\lambda = 450$ and 650 nm at 2 GeV. Adapted from [94]

when electrons approach a metallic surface. This radiation has a characteristic polarisation, spectrum and intensity. A variation to this radiation occurs when the charged particle moves roughly parallel to a conducting undulated surface. An oscillating dipole will be set up with a frequency related to the particle velocity and the undulation. The radiated power is small, but due to the microscopic source area, the brightness can be large. This has, amongst a range of other usages, found an application in accurate beam diagnostics equipments.

As an example, we will use an experiment to investigate the geometrical resolution of optical transition radiation as shown in Fig. 7.31a [94]. Integrating Eq. (7.40) over the solid angle gives:

$$\frac{dN}{d\omega} \simeq \frac{2\alpha}{\pi\omega} \ln(\gamma\Theta_{\max}) \tag{7.48}$$

where $\Theta_{\max}$ is the angle of maximum emission, measured by the optical spectrometer. The number of photons emitted is small. This must be compensated by a large number of particles in the beam.

The mathematics for such a set-up is given in [95]. The diffraction, or the Heisenberg uncertainty principle in the transverse phase-space of the photon, sets the lower limit for the size of the emitting surface:

$$\Delta b_i \geq \frac{\lambda}{2\pi} \frac{1}{2\Delta\Theta_i} \quad (i = x, y) \tag{7.49}$$

where $\lambda \sim 600$ nm is the observed wavelength. b_i and Θ_i are the components of the impact parameter **b** and the photon direction. $\Delta\Theta_i$ and Δb_i refer to rms values. Setting $\Delta\Theta = \gamma^{-1}$, or full acceptance for the photons, the resolution becomes

Table 7.7 Parameters for the fit to the data [94] and plotted in Fig. 7.31b

Parameter	$\lambda = 450\,\text{nm}$	$\lambda = 650\,\text{nm}$
ρ	$176 \pm 12\,\mu\text{m}$	$163 \pm 25\,\mu\text{m}$
a	$(9 \pm 5) \cdot 10^{-5}$	$(6 \pm 3) \cdot 10^{-5}$
b	1.12 ± 0.09	1.12 ± 0.06

proportional to γ. $\gamma = 10^5$ would give $\Delta b \geq 5\,\text{mm}$. This effect can be limited by the introduction of an iris in the optical path as in [96].

The results from [94] are shown in Fig. 7.31b. As expected, the resolution is weakly dependent on the intensity of the beam, but the total uncertainty is small. The measurement points are fitted to $\sigma_{\text{rms}} = \sqrt{\rho^2 + aI^b}$, where a and b are fit parameters, ρ is the real beam dimension and I is the beam intensity. These are given in Table 7.7.

Another promising application for optical transition radiation is in aerogel[23] Cherenkov detectors [97].

7.5.3.2 X-ray Transition Radiation Detectors

Following [98], the total radiated energy from a single surface per unit of frequency, can be approximated by:

$$\left[\frac{dW}{d\omega}\right]_{\text{s.s.}} = \frac{\alpha}{\pi}\left[\frac{1 + r + 2X_1^2}{1 - r}\ln\frac{X_1^2 + 1}{X_1^2 + r} - 2\right] \tag{7.50}$$

where

$$X_1 = \frac{\omega}{\gamma\omega_{p1}} \quad \text{and} \quad r = \frac{\omega_{p2}^2}{\omega_{p1}^2} \sim \frac{\rho_2}{\rho_1} \tag{7.51}$$

The suffix $_1$ and $_2$ denote medium 1 and 2. ω_{pi} is the plasma frequency for medium i. r will be assumed to be small and in the range of 10^{-3}, which corresponds to a $\rho = 1$ to gas interface.

By analysing Eq. (7.50), which is plotted in Fig. 7.32a, three distinct regimes can be examined:

1. If $\gamma \ll \omega/\omega_{p1}$ then $X_1 \gg 1$ and $dW/d\omega \sim \alpha/6\pi X_1^4$, which is a small number. This results in a frequency cut-off and thereby $\omega \leq \gamma\omega_{p1}$.
2. If $\omega/\omega_{p1} \ll \gamma \ll \omega/\omega_{p2}$ then $dW/d\omega \propto \ln X_1^{-1}$. That is, the total radiated power increases logarithmically with γ.
3. If $\gamma \gg \omega/\omega_{p2}$ then $X_1 \ll \sqrt{r}$. Then the total radiated power is approximately constant.

[23] See Sect. 7.4.2.2.

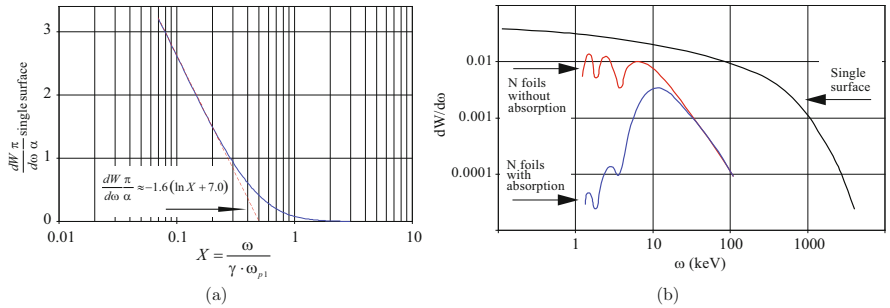


Fig. 7.32 (**a**) Total radiated energy from a single surface per unit of frequency as function of the dimensionless variable $X = \omega/\gamma\omega_{p1}$. (**b**) Intensity of the forward radiation divided by the number of interfaces for 20 μm polypropylene ($\omega_p = 21$ eV) and 180 μm helium ($\omega_p = 0.27$ eV). Adapted from [99]

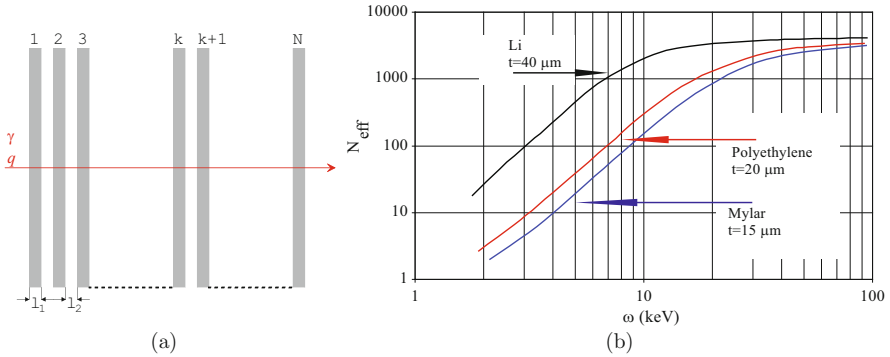


Fig. 7.33 (**a**) Sketch of a periodic transition radiation radiator. (**b**) The effective number of foils in a radiator as function of photon energy. Adapted from [90]

It can be shown that the mean radiated energy in this single surface configuration can be written as:

$$W \simeq 2\alpha\gamma\omega_{p1}/3 \tag{7.52}$$

and that the number of high energy photons produced are of the order of α when taking into account the frequency cut-off discussed above:

$$N_{\text{photons}}(\omega > 0.15\gamma\omega_{p1}) \simeq \alpha/2 \tag{7.53}$$

A large number of interfaces are therefore required to have an effective detector with a sufficient signal-to-noise ratio. A periodic transition radiation radiator is sketched in Fig. 7.33a. It should be noted that the radiators do not need to be rigorously periodic, but it is helpful for the calculation of the yield.

The basic mathematics can be found in [85, 90, 98]. Computational models can be found in [100]. The effective final number of transition radiation high energy photons at the end of the radiator stack is a function of constructive and destructive effects. See Fig. 7.32b. We will list the main effects here:

- The total radiated energy of a single surface is proportional to the plasma frequency and thereby proportional to $\sqrt{Z}$ of the material. Equation (7.44). The absorption of these photons is governed by photo-electric effects and the absorption coefficients in the stack. This goes approximately like Z^5. The radiator material should therefore be of low Z.
- The thickness of the radiator material, $\mathtt{l}_1$ in Fig. 7.33a, must be large enough to contain the formation zone for the required γ, but short enough not to introduce multiple scattering effects and bremsstrahlung. The gas density will always introduce a negative effect and should be kept as low as possible.

For a practical transition radiation radiator and following [90], the expression of the total flux, is then represented by an integration over the emission angle Θ and a function which represents the incoherent addition of the single foil intensities and includes the photon absorption in the radiator. The effective number of foils in the radiator can then be expressed as:

$$N_{\text{eff}} \simeq \frac{1 - \exp(-N\sigma)}{1 - \exp(-\sigma)} \tag{7.54}$$

where $\sigma = (\kappa\rho t)_{\text{foil}} + (\kappa\rho t)_{\text{gas}}$ and κ, ρ and t are respectively the absorption coefficient, density and thickness of the material. The self-absorption of the photons from transition radiation limits the yield and $N_{\text{eff}} \to 1/\left[1 - \exp(-\sigma)\right]$ for $N \to \infty$. A typical mean energy for the photons in a practical radiator is in the range of 10 keV. See Fig. 7.32b. The spectrum will be softer for foils with lower plasma frequencies. Since N_{eff} in Eq. (7.54) is depending on the absorption coefficient through the frequency of the photon, N_{eff} will saturate for high frequencies as shown in Fig. 7.33b.

7.5.3.3 X-ray Detectors

Any detector which has a sufficiently high efficiency for X-rays of the order of 10 keV can be used. In the design of the detector it should be noted that the number of transition radiation photons is small and produced very close to the path of the charged particle which will normally also traverse the detector. The traditional detector is a MWPC-like detector, Chap. 3, which directly follows the radiator. In order to enhance the signal-to-noise ratio and efficiently use the space as the effective number of interfaces in the radiator will saturate, a transition radiation detector is therefore normally many radiator/detector assemblies.

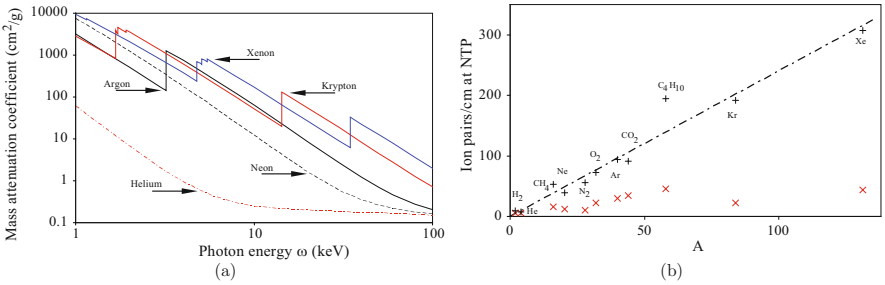


Fig. 7.34 (**a**) X-ray mass attenuation coefficient, μ/ρ, as function of the photon energy. $\mu/\rho = \sigma_{tot}/uA$, where $u = 1.660 \times 10^{-24}$ g is the atomic mass unit, A is the relative atomic mass of the target element and σ_{tot} is the total cross section for an interaction by the photon. Data from http://physics.nist.gov/PhysRefData/. (**b**) The (×) primary and (+) total number of ion pairs created for a minimum ionizing particle per cm gas at normal temperature and pressure as function of molecular mass A [101]

The ionization loss, dE/dx, from the charged particle will create charge clusters. Some of them rather far from the track due to δ-electrons. The absorption of transition radiation photons will produce a few local strong charge clusters. The choice of gas is therefore a compromise between photon absorption length, Fig. 7.34a, and the background from dE/dx, Fig. 7.34b. The optimal gas thickness is about one absorption length for 10 keV. Xenon is the preferred gas with a chamber thickness of 10–15 mm. See discussion in [90]. CO_2, or similar, is added as quencher.

A minimum ionizing particle, MIP, will produce a total of ∼310 ion pairs per cm xenon gas. Figure 7.34b. The relativistic rise is about 75% in xenon at 1 atm, or about 550 ion pairs/cm will be produced by a high γ charged particle. The average energy required to create an ion pair in a gas, is typically 25–35 eV. For xenon it is measured to 22.1 ± 0.1 eV [102], or about double the ionization energy for the least tightly bound shell electron. A 10 keV transition radiation photon will then produce about 450 ion pairs. The signal-to-noise ratio will be further reduced due to Landau-fluctuations and gain variations in the detector and electronics. Additional background might arise from curling in a magnetic field, bremsstrahlung and particle conversions. The challenge is then to correctly identify the photon cluster from a dE/dx signal of about the same strength. We will illustrate this by looking more closely at the choices made by the ALICE [103] and ATLAS [104] experiments.

7.5.3.4 ATLAS Transition Radiation Tracker

In the ATLAS experiment, the transition radiation tracker (TRT) in the barrel comprises many layers of gaseous straw tube elements interleaved with transition radiation material. Figure 7.35. With an average of 36 hits per track, it provides continuous tracking to enhance the pattern recognition and improve the momentum

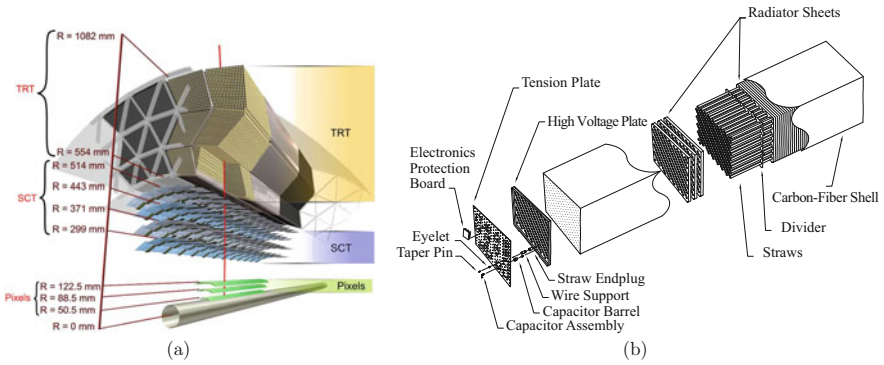


Fig. 7.35 (**a**) ATLAS Detector. Drawing showing the sensors and structural elements traversed by a charged track of 10 GeV p_t in the barrel inner detector (pseudo rapidity $\eta = 0.3$). The track traverses approximately 36 axial straws of 4 mm diameter contained in the barrel transition-radiation tracker modules. [104]. (**b**) Layout of an ATLAS Barrel TRT module. The ATLAS TRT collaboration et al. [105] with permission

resolution over $|\ \eta\ | < 2.0$[24] and electron identification complementary to that of the calorimeter over a wide range of energies. A similar detector is placed in the forward direction.

The transition radiator material which completely surrounds the straws inside each module, Fig. 7.35b, consists of polypropylene-polyethylene fibre mat about 3 mm thick. The fibres are typically 19 μm in diameter and are formed from polyethylene clad polypropylene material. The fibres are formed into fabric plies with 3 mm thickness and a density of about 0.06 g/cm^3. The absorption length for the lowest energy photons of interest (5 keV) is about 17 mm in the radiator material.

The ATLAS TRT uses two thresholds to discriminate between digitisations from tracks and those from transition radiation:

1. Low threshold, LT, for tracking which is set to ∼300 eV with 8 digitisations over 25 ns.
2. High threshold, HT, set in the range 5–7 keV with 1 digitisation over 25 ns and read out in 75 ns segments.

As the $\beta\gamma$ of the traversing particles will vary greatly, and thereby the ionization in the straw tubes, a Time-over-Threshold parameter can be defined from the LT digitisations in order to enhance the signal-to-noise estimate for the transition radiation signal.

Particle identification properties of the TRT Barrel using transition radiation were studied at several different beam energies. The good agreement between 2 GeV low

[24] Pseudo rapidity, η, is describing the angle of a particle relative to the beam axis. $\eta = -\ln\left[\tan\left(\frac{\Theta}{2}\right)\right] = \frac{1}{2}\ln\left[\frac{|\mathbf{p}|+p_L}{|\mathbf{p}|-p_L}\right]$. Θ is the angle between the particle momentum and the beam axis, $\mathbf{p}$ is the momentum vector and p_L is the longitudinal momentum component.

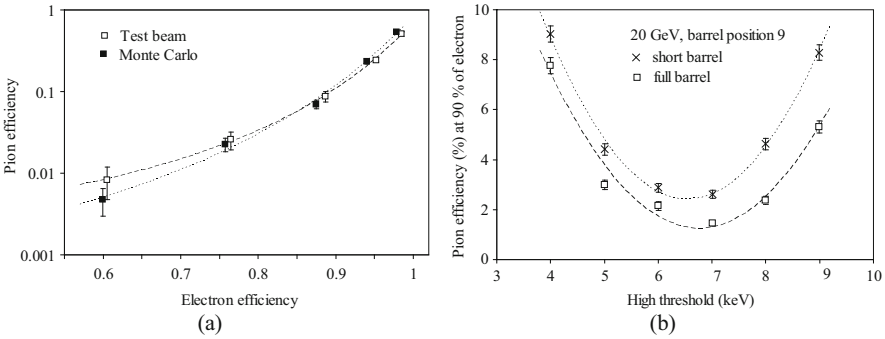


Fig. 7.36 (**a**) ATLAS TRT test beam. Pion rejection curve for a 2 GeV e/π beam. Cornelissen and Liebig [106] with permission. (**b**) ATLAS TRT test beam. e/π rejection power as a function of the high level threshold. Full barrel: all barrel straw layers are active. Short barrel: particle crosses the barrel in the central area where the first 9 layers do not have active anode wires. The ATLAS TRT collaboration et al. [105] with permission

energy data and simulation is shown in Fig. 7.36a. The results for 20 GeV beam energy are shown in Fig. 7.36b. On this figure the pion rejection power is shown as a function of the high level threshold at two beam positions along the straw. The upper points are when beam particles crossed the Barrel module 40 cm from its edge. At this position the first 9 straw layers are not active. The lower points are when the beam is positioned 20 cm from the edge of the Barrel where all 73 straw layers are active. As seen in this figure the best particle identification properties for the TRT Barrel are at a threshold of about 7 keV. Pion mis-identification in that case is 1.5–3% at 90% of the electron efficiency.

7.5.3.5 ALICE Transition Radiation Detector

The main purpose of the ALICE Transition Radiation Detector (TRD) [103, 107] is to provide electron identification in the central barrel for momenta above 1 GeV/c. Below this momentum electrons can be identified via specific energy loss measurement in the TPC. Above 1 GeV/c transition radiation from electrons passing a radiator can be exploited together with the specific energy loss in a suitable gas mixture to obtain the necessary pion rejection capability. The chamber geometry and the read-out electronics were chosen to reconstruct track segments. Since the angle of the track segment with respect to the origin is a measure of the transverse momentum of the electron, this information is used in the first level trigger within 5 μs of the collision.

The pion rejection is governed by the signal-to-background ratio in the measurement of J/Ψ production and its p_t dependence. This led to the design goal for the pion rejection capability of a factor 100 for momenta above 1 GeV/c in central Pb-Pb collisions.

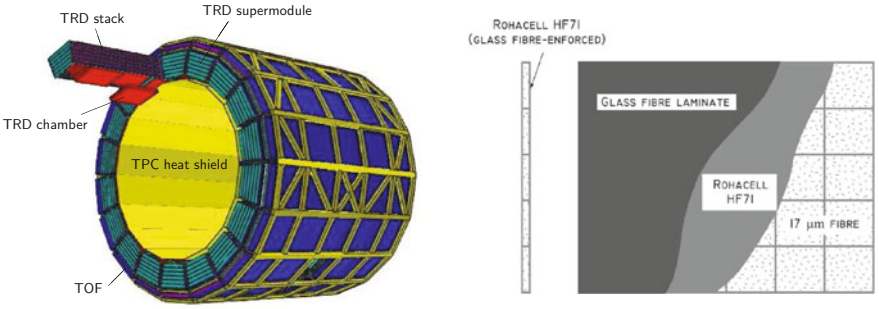


Fig. 7.37 (**a**) Schematic drawing of the TRD layout in the ALICE space frame. Shown are 18 super modules each containing 30 readout chambers (red) arranged in five stacks of six layers. One chamber has been displaced for clarity. On the outside the TRD is surrounded by the Time-Of-Flight (TOF) system (dark blue). On the inside the heat shield (yellow) towards the TPC is shown. The ALICE Collaboration et al. [103] with permission. (**b**) The principle design of the TRD sandwich radiator. The ALICE Collaboration et al. [107] with permission

The TRD consists of 540 individual readout detector modules. Figure 7.37a. Each detector element consists of a carbon fibre laminated Rohacell[25]/polypropylene fibre sandwich radiator, Fig. 7.37b, of 48 mm thickness, a drift section of 30 mm thickness, or about 2 µs, and a multi-wire proportional chamber section (7 mm) with pad readout.

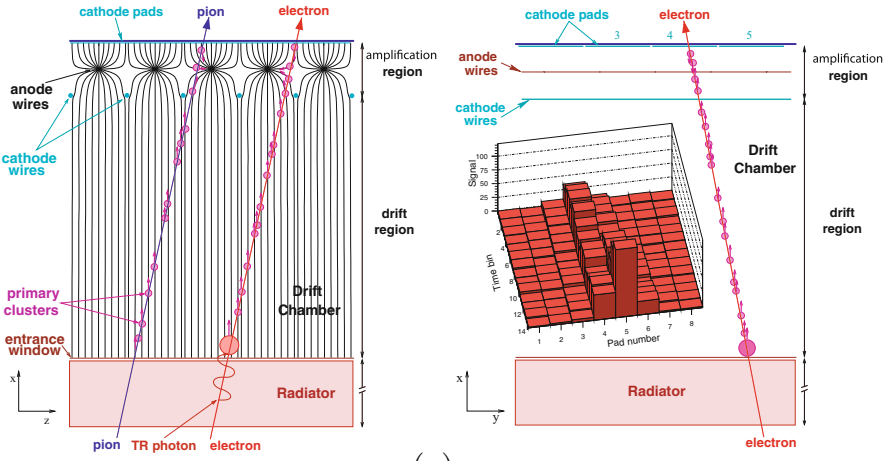
Following [108], employing the drift time information in a bidimensional likelihood [109], the pion rejection capability can be improved by about 60% [110] compared to the standard likelihood method on total deposited charge. This method is the simplest way of extending the standard method. However, it does not exploit all recorded information, namely the amplitude of the signal in each time bin. Along a single particle track this information is highly correlated, Fig. 7.38a, due to

- the intrinsic detector signal, in particular since a Xe-based mixture is used
- the response of the front-end electronics used to amplify the signals.

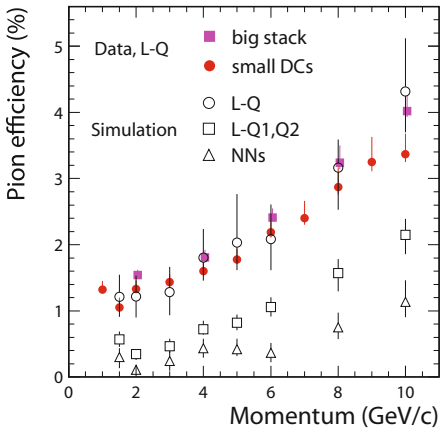
Under these circumstances, the usage of a neural network (NN) algorithm is a natural choice for the analysis of the data. The result of the data analysis from a 2–6 GeV/c mixed e/π test beam is shown in Fig. 7.38b [108]. Neural Network algorithm might improve the pion rejection significantly by a factor larger than 3 for a momentum of 2 GeV/c compared to other methods.

The detector was completed in the LS 1 before RUN 2 at LHC. Since then it provides coverage of the full azimuthal acceptance of the central barrel. Figure 7.39 shows the p_T spectra of electron candidates with 6 layers identified using the TPC and the TOF in the minimum-bias and triggered data sample. The expected onset

[25] ROHACELL is a close cell polymethacrylimide- (PMI-) rigid foam by Evonik Industries AG, Germany.



(a)



(b)

Fig. 7.38 (**a**) Schematic cross-sectional view of an ALICE detector module in rz and $r\phi$-direction. The inset shows the charge deposit from an inclined track which is used for momentum reconstruction. The ALICE Collaboration et al. [103] with permission. (**b**) Measured pion efficiency as a function of beam momentum applying likelihood on total deposited charge (L-Q) (full symbols) measured with a stack of six chambers and smaller test chambers. Results are compared to simulations (open symbols) for 90% electron efficiency and six layers. These simulations were extended to two-dimensional likelihood on deposited charge and position (LQ1, Q2) and neural networks (NN). The ALICE Collaboration et al. [103] with permission

at the trigger threshold of 3 GeV/c is observed for the triggered events and shows in comparison to the corresponding spectrum from minimum-bias collisions an enhancement of about 700. At 90% electron efficiency, a pion rejection factor of

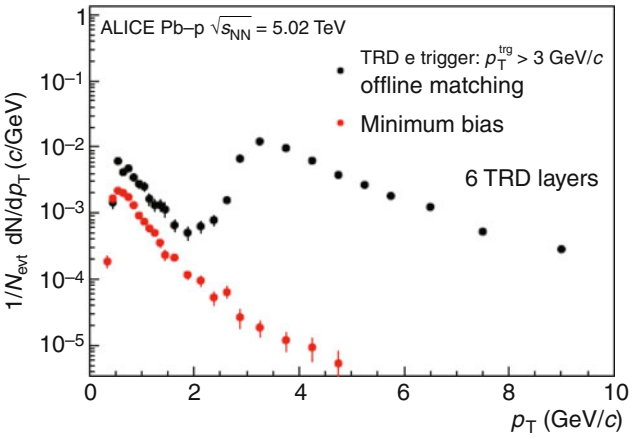


Fig. 7.39 p_T spectra of identified electrons for the minimum-bias and TRD-triggered data sample of Pb-p collisions at $\sqrt{s_{NN}} = 5.02$ TeV. For the result of the TRD-triggered sample, electrons from photon conversions in the detector material were rejected by matching the online track with a track in the TPC. Reference [111]

about 70 is achieved at a momentum of 1 GeV/c for simple identification algorithms. When using the temporal evolution of the signal, a pion rejection factor of up to 410 is obtained.

References

1. B. Dolgoshein, Complementary particle ID: transition radiation and dE/dx relativistic rise, Nucl. Instrum. Methods Phys. Res., A: 433 (1999) 533
2. C. Ward et al., A Large Aperture Time-Of-Flight Counter System, Nucl. Instr. and Meth. 30(1964)61–68
3. G. Charpak, L. Dick and L. Feuvrais, Location of the position of a particle trajectory in a scintillator, Nucl. Instr. and Meth. 15(1962)323–326
4. W. Klempt, Review of particle identification by time of flight techniques, Nucl. Instrum. Methods Phys. Res., A: 433 (1999) 542
5. A. Rabl and R. Winston, Ideal concentrators for finite sources and restricted exit angles, Applied Optics, Vol. 15 Issue 11, pp.2880–2883 (1976)
   R.H. Hildebrand and R. Winston, Throughput of diffraction-limited field optics systems for infrared and millimetric telescopes: erratum, Applied Optics, Vol. 21 Issue 17, pp.3065–3065 (1982)
6. W. Blaschke and G. Thomsen, Vorlesungen Uber Differential-Geometrie und Geometrische Grundlagen von Einsteins Relativitatstheorie. I: Elementaire Differentialgeometrie, Dritte Erweiterte Auflage. Bearbeitet under herausgegeben, DOVER PUBLICATIONS (1945)
7. T. Kobayashi and T. Sugitate, Test of Prototypes for a Highly Segmented TOF Hodoscope, Nucl. Instrum. Methods Phys. Res., A: 287 (1990) 389–396
8. C. Field et al., Development of photon detectors for a fast focusing DIRC, Nucl. Instrum. Methods Phys. Res., A 553 (2005) 96–106

9. J. Blazej et al., Gallium-based avalanche photodiode optical crosstalk, Nucl. Instrum. Methods Phys. Res., A: 567 (2006) 239–241, and references therein.
10. J. Va'vra et al., A 30 ps timing resolution for single photons with multi-pixel Burle MCP-PMT, Nucl. Instrum. Methods Phys. Res., A: 572 (2007) 459–462
11. A.N. Akindinov et al., Latest results on the performance of the multigap resistive plate chamber used in the ALICE TOF, Nucl. Instrum. Methods Phys. Res., A: 533 (2004) 74–78
12. J. Va'vra, PID techniques: Alternatives to RICH methodes, Nuclear Instruments & Methods in Physics Research A (2017), http://dx.doi.org/10.1016/j.nima.2017.02.075
13. A. Bravar, Recent Results from NA61 (Flux Related Systematic Uncertainties) and Recent Results from T2K (Overall Systematic Uncertainties), http://www.t2k.org/docs/talk/225, 2015-11-06
N. Abgrall et al., Measurements of cross sections and charged pion spectra in proton-carbon interactions at 31 GeV/c, PHYSICAL REVIEW C 84, 034604 (2011)
14. S. Afanasiev et al., The NA49 Large Acceptance Hadron Detector, CERN-EP/99-001, 6. January 1999
15. P.A. Cherenkov, Visible glow under exposure of gamma irradiation, Dokl. Acad. Nauk. USSR 2(1934)451
S.I. Vavilov, About possible reasons of blue gamma-glow of liquids, Dokl. Acad. Nauk. USSR 2(1934)457
I.M. Frank and I.E. Tamm, Coherent Radiation of Fast Electrons in a Medium, Dokl. Acad. Nauk. USSR 14(1937)107
16. J.V. Jelley, Čerenkov Radiation and its Application, London: Pergamon Press, 1958
V.P. Zrelov, Čerenkov Radiation in High Energy Physics, Jerusalem, Israel Program for Scientific Translations, 1970
17. J. Séguinot, Les Compteurs Cherenkov: Applications et Limites pour l'Identification des Particules. Développements et Perspectives, CERN-EP/89–92, LPC/89-25, 31 juillet 1989
18. J.D. Jackson, Classical Electrodynamics, John Wiley & Sons, Inc.,1998,
ISBN 0-471-30932-X
19. W.W.M. Allison and P.R.S. Wright, The Physics of Charged Particle Identification, in:
T. Ferbel (ed.), Experimental Techniques in High-Energy Nuclear and Particle Physics, ISBN 981-02-0867-7, World Scientific Publishing Company, Inc. 1999.
20. Landolt-Börnstein, Eigenschaften der Materie in ihren Aggregatzuständen, 8. Teil Opische Konstanten.
21. C.C. Chen and W.A. Steel, Molecular calculation of the dielectric constant of N_2 and CO_2, J. Chem. Phys., 75(1)383–387
22. DuPont, Freon Technical Bulletin B-32, 32A
23. P.W. Langhoff and M. Karplus, Pade Summation of the Cauchy Equation, Journal of the Optical Society of America, 59(1969)863
24. J. Koch, Nova acta Soc. Upsal. (4) Nr. 5, 1909
C. u. M. Cuthbertson, Proc. Roy. Soc. London (A) 83, 151, 1910
25. E. Albrecht et al., VUV absorbing vapours in n-perfluorocarbons, Nucl. Instr. and Meth. in Phys. Res. A 510(2003)262–272
26. D.E. Gray (ed.), American Institute of Physics handbook, McGraw-Hill Book Company. Inc., 1957
27. E. Efimov and S. Stone, A novel LiF radiator for RICH detectors, Nucl. Instr. and Meth. in Phys. Res. A 371(1996)79–81
28. B. Dey et al., Design and performance of the focusing DIRC detector, Nuclear Instruments and Methods in Physics Research A 775(2015)112–131
J. Schwiening, The DIRC detector at the SLAC B-factory PEP-II: operational experience and performance for physics application, Nucl. Instr. and Meth. in Phys. Res. A 502(2003)67–75
P. Coyle et al., The DIRC counter: a new type of particle identification device for B factories, Nucl. Instr. and Meth. in Phys. Res. A 343(1994)292–299

29. Blair N. Ratcliff, Imaging rings in Ring Imaging Cherenkov counters, Nucl. Instr. and Meth. in Phys. Res. A 502(2003)211–221
30. J. Fast, The Belle II imaging Time-of-Propagation (iTOP) detector, Nuclear Instruments & Methods in Physics Research A (2017), http://dx.doi.org/10.1016/j.nima.2017.02.045
Y. Enari et al., Progress report on Time-Of-Propagation counter—a new type of ring imaging Cherenkov detector, Nucl. Instr. and Meth. in Phys. Res. A 494(2002)430–435
31. G. S. Varner, Waveform-sampling electronics for Cherenkov detectors, https://indico.cern.ch/event/393078/contributions/2241768/attachments/1334816/2007370/1-WaveformReadout_Varner_RICH2016.pdf
32. M.J. Charles and R. Forty, TORCH: Time of flight identification with Cherenkov radiation, Nuclear Instruments and Methods in Physics Research A 639(2011)173?176
T. Gys et al., The TORCH detector R & D: Status and perspectives, Nuclear Instruments and Methods in Physics Research A (2017), http://dx.doi.org/10.1016/j.nima.2017.02.060
33. C. Schwarz et al., The PANDA DIRC Detectors at FAIR, arXiv:1707.09269 [physics.ins-det]
34. M. Cantin et al., Silica aerogels used as Cherenkov radiators, Nucl. Instr. and Meth. 118(1974)177–182
35. H. Gordon et al., The Axial Field Spectrometer at the CERN ISR, CERN-EP/81–34
36. L. A. Paquette (ed), D. Crich (ed), Ph. L. Fuchs (ed) and G. Molander (ed), Encyclopedia of Reagents for Organic Synthesis, Wiley; 2 edition (May 18, 2009)
L. Rösch, P. John and R. Reitmeier, Silicon Compounds, Organic, Ullmann's Encyclopedia of Industrial Chemistry, Wiley-VCH; Sixth edition (February 21, 2003)
37. T. Bellunato et al., Refractive index of aerogel: uniformity and dispersion law, Nucl. Instr. and Meth. in Phys. Res. A 595(2008)183–186
T. Bellunato et al., Refractive index dispersion law of silica aerogel, Eur. Phys. J. C52(2007)759–764
38. D. Perego. Private communication, March 2008
39. A.Yu. Barnyakov et al., Focusing Aerogel RICH optimization, Nucl. Instr. and Meth. in Phys. Res. A 595(2008)100–103
40. M.Born and E. Wolf, Principles of Optics, University Press, Oxford, ISBN 0-521-64222-1
41. R. Pestotnik et al., The aerogel Ring Imaging Cherenkov system at the Belle II spectrometer, Nuclear Instruments & Methods in Physics Research A (2017), http://dx.doi.org/10.1016/j.nima.2017.04.043
42. O. Chamberlain, E. Segre, C. Wiegand and Th. Ypsilantis, Observation of antiprotons, Phys. Rev. 100, No. 3, 947 (1955).
43. J. Litt and R. Meunier, Čerenkov counter technique in high-energy physics, Annu. Rev. Nucl. Part. Sci.: 23(1973)1–43
R. Meunier, Čerenkov Detectors, 1973 International Conference on Instrumentation for High Energy Physics, Frascati, Italy, May 8–12, 1973, Published by Laboratori Nazionali del CNEN
44. M. Bourquin et al., Particle and antiparticle production by 200 GeV/c protons in the charged hyperon beam at the CERN SPS, Nucl. Phys. B 153(1979)13–38
45. J. Séguinot and T. Ypsilantis, Photoionization and Cherenkov Ring Imaging, Nucl. Instr. and Meth., 142(1977)377; for an extensive list of RICH references, see http://alice-hmpid.web.cern.ch/alice-hmpid/basic_references.htm
46. J. Séguinot and T. Ypsilantis, A historical survey of ring imaging Cherenkov counters, Nucl. Instr. and Meth. in Phys. Res. A, 343(1994)1–29
J. Séguinot and T. Ypsilantis, Theory of ring imaging Cherenkov counters, Nucl. Instr. and Meth. in Phys. Res. A, 343(1994)30–51
47. Carl A. Heller and Aaron N. Fletcher, Oxidation and Chemiluminescence of Tetrakis (dimethylamino) ethylene. I. Reversible Reactions of Oxygen with Tetrakis (dimethylamino) ethylene and n-Decane, J. Phys. Chem.; 1965; 69(10); 3313–3317
J.P. Paris, Chemiluminescence of tetrakis-(dimethylamino)-ethylene, Photochemistry and Photobiology, 4(6)(1965)1059–1065
Sidney Toby, Paul A. Astheimer and Frina S. Toby, Chemiluminescence in the gas phase

reaction between tetrakis-(dimethylamino)-ethylene and oxygen, Journal of Photochemistry and Photobiology A: Chemistry 67(1992)1–12
W.R. Carpenter and E. M. Bens, Influence of oxidation products on the chemiluminescent oxidation of tetrakis-(dimethyl amino)-ethylene. Naval ordnance test station China Lake CA, Report 0526608, FEB 1967
W.P. Norris, Reactions of Tetrakis-(Dimethylamino)-Ethylene with Weak Acids. Naval ordnance test station China Lake CA, Report 0111409, OCT 1972
48. D. F. Anderson, A xenon gas scintillation proportional counter coupled to a photoionization detector, Nucl. Instr. and Meth., 178(1980)125–130
Wiliam H. -M. Ku, Charles J. Hailey and Michael H. Vartanian, Properties of an imaging gas scintillation proportional counter, Nucl. Instr. and Meth., 196(1982)63–67
D. F. Anderson, Measurement of TMAE and TEA vapor pressures, Nucl. Instr. and Meth. in Phys. Res. A, 270(1988)416–418
Richard A. Holroyd et al., Measurement of the absorption length and absolute quantum efficiency of TMAE and TEA from threshold to 120 nm, Nucl. Instr. and Meth. in Phys. Res. A, 261(1987)440–444
49. H.W. Siebert et al., The Omega RICH, Nucl. Instr. and Meth. in Phys. Res. A, 343(1994)60–67
50. D. Aston et al., Development and construction of the SLD Cherenkov ring-imaging detector, Nucl. Instr. and Meth. in Phys. Res. A, 283(1989)582–589
51. E. Albrecht et al., Operation, optimisation, and performance of the DELPHI RICH detectors, Nucl. Instr. and Meth. in Phys. Res. A, 433(1999)47–58
52. M. Starič, A. Stanovnik and P. Križan, Tests of a solid CsI photocathode in a thin multiwire proportional chamber, Nucl. Instr. and Meth. in Phys. Res. A, 307(1991)145–148
J. Almeida et al., Development of large area fast-RICH prototypes with pad readout and solid photocathodes, Nucl. Instr. and Meth. in Phys. Res. A. 348(1994)216–222
53. J.R. Hardy and A.M. Karo, The Lattice Dynamics and statics of alkali halide crystals, Plenum Press 1979, ISBN 0-306-4022-1
I.P. Csorba, Imaging Tubes, Howard W, Sams & Co., Inc. 1985, ISBN 0-672-22023-7
54. C. Lu and K. T. McDonald, Properties of reflective and semitransparent CsI photocathodes, Nucl. Instr. and Meth. in Phys. Res. A. 343(1994)135–151
A. Braem et al., Technology of photocathode production, Nucl. Instr. and Meth. in Phys. Res. A, 502(2003)205–210
55. T. Meinschad, L. Ropelewski and F. Sauli, GEM-based photon detector for RICH applications, Nucl. Instr. and Meth. in Phys. Res. A. 535(2004)324–329
A. Breskin et al., Recent advances in gaseous imaging photomultipliers, Nucl. Instr. and Meth. in Phys. Res. A. 513(2003)250–255
J.M. Maia et al., Single-UV-photon 2-D imaging with multi-GEM detectors, Nucl. Instr. and Meth. in Phys. Res. A. 580(2007)373–376
L. Periale et al., Detection of the primary scintillation light from dense Ar, Kr and Xe with novel photosensitive gaseous detectors, Nucl. Instr. and Meth. in Phys. Res. A. 478(2002)377–383, and references therein.
56. F. Tessarotto, Status and perspectives of gaseous photon detectors, Nuclear Instruments & Methods in Physics Research A (2017), http://dx.doi.org/10.1016/j.nima.2017.023.011
57. I. Ariño et al., The HERA-B RICH, Nucl. Instr. and Meth. A 453 (2000) 289–295
58. P. Abbon et al., Read-out electronics for fast photon detection with COMPASS RICH-1, Nucl. Instr. and Meth. A 587 (2008) 371–387
59. M. Moritz et al., Performance Study of New Pixel Hybrid Photon Detector Prototypes for the LHCb RICH Counters, IEEE TRANSACTIONS ON NUCLEAR SCIENCE, 51(2004)1060–1066
60. LHCb Collaboration, LHCb RICH, Technical Design Report, CERN/LHCC/2000-0037, 7 September 2000
61. A. Braem et al., Development, fabrication and test of a highly segmented hybrid photodiode, Nucl. Instr. and Meth. in Phys. Res. A. 478(2002)400–403

62. T. Gys, Micro-channel plates and vacuum detectors, Nuclear Instruments & Methods in Physics Research A 787(2015)254–260
T. Iijima, Status and perspectives of vacuum-based photon detectors, Nuclear Instruments & Methods in Physics Research A 639(2011)137–143
T. Gys, Status and perspectives of vacuum-based photon detectors for single photon detection, Nuclear Instruments & Methods in Physics Research A 595(2008)136–141
63. ALICE Collaboration, ALICE Technical Design Report, Detector for High Momentum PID, CERN/LHCC 98–19, ALICE TDR 1, 14 August 1998.
64. J. Holder et al., The first VERITAS telescope, Astroparticle Physics 25 (2006) 391–401
65. P. Baillon et al., An improved method for manufacturing accurate and cheap glass parabolic mirrors, Nucl. Instrum. Methods Phys. Res., A 276(1989)492–495
P. Majumdar et al., Angular Resolution of the Pachmarhi Array of Cerenkov, Telescopes Astropart. Phys. 18(2003)333–349
66. M. Laub, Development of opto-mechanical tools and procedures for the new generation of RICH-detectors at CERN, CERN-THESIS-2006-028; LHCb-2001-130; CERN-LHCb-2001-130, Prague: Prague TU, 2001
67. E. Albrecht et al., The mirror system of COMPASS RICH-1, Nucl. Instrum. Methods Phys. Res., A 502 (2003) 236–240
68. G.J. Barber et al., Glass-coated beryllium mirrors for the LHCb RICH1 detector, Nucl. Instrum. Methods Phys. Res., A 570(2007)565–57
69. J. Friese, R. Gernhäuser, P. Maier-Komor and S. Winkler, A new carbon based VUV mirror of high radiation length for the HADES RICH, Nucl. Instrum. Methods Phys. Res., A 502(2003)241–245
F.C.D. Metlica and On behalf of the LHCb Collaboration, Development of light-weight spherical mirrors for RICH detectors, Nucl. Instrum. Methods Phys. Res., A 595(2008)197–199
70. T. Bellunato et al., Light composite mirrors for RICH detectors: production, characterisation and stability tests, Nucl. Instrum. Methods Phys. Res., A 538(2005)458–464
71. C. Bigongiari et al., The MAGIC telescope reflecting surface, Nucl. Instrum. Methods Phys. Res., A 518(2004)193–194
72. U. Müller et al., The Omega RICH in the CERN hyperon beam experiment, Nucl. Instrum. Methods Phys. Res., A 433(1999)71–76
73. A. Braem, C. David and C. Joram, Metal multi-dielectric mirror coatings for Cherenkov detectors, Nucl. Instrum. Methods Phys. Res., A 553(2005)182–186
74. D. Malacara (Ed.), Optical Shop Testing (Wiley Series in Pure and Applied Optics), Wiley-Interscience, 2007, ISBN: 978-1574446821
D. Malacara, M. Servín and Z. Malacara, Interferogram Analysis For Optical Testing, Second Edition (Optical Engineering), CRC; 2005, ISBN: 978-0471484042
75. The LHCb Collaboration, A. Augusto Alves Jr. et al, The LHCb Detector at the LHC, JINST 3 (2008) S08005.
76. G. Wilkinson, In search of the rings: Approaches to Cherenkov ring finding and reconstruction in high-energy physics, Nucl. Instr. and Meth. A 595 (2008) 228.
77. M. Adinolfi et al., Performance of the LHCb RICH detector at the LHC, Eur. Phys. J. C73 (2013) 2431, arXiv:1211.6759.
78. R. Forty and O. Schneider, RICH Pattern Recognition, LHCb note LHCb-98-040.
79. C. Buszello, LHCb RICH pattern recognition and particle identification performance, Nucl. Instr. and Meth. A 595 (2008) 245.
80. M. Starič, Track based maximum likelihood ring search, Nucl. Instr. and Meth. A 595 (2008) 237
81. LHCb collaboration: R. Aaij et al., Measurement of b-hadron branching fractions for two-body decays into charmless charged hadrons, JHEP10(2012)037
82. E. Fermi, The Ionization Loss of Energy in Gases and in Condensed materials, Phys. Rev. 57(1940)485–493

83. V.L. Ginzburg and I.M. Frank, Radiation of a uniformly moving electron due to its transition from one medium into another, Zh. Eksp. Teor. Fiz. 16 (1946)15.
a shortened (English) version J. Phys. USSR 9 (1945), 353.
84. C.W. Fabjan and W. Struczinski, Coherent Emission of Transition Radiation in Periodic Radiators, Physics Letters 57B(1975)483–486
85. C. Leroy and P-G. Rancoita, Principles of radiation interaction in matter and detection, World Scientific, 2004, ISBN: 978-9812389091
86. M.L.Ter-Mikaelian, High-Energy Electromagnetic Processes in Condensed Media, John Wiley & Sons, Inc, 1972, ISBN 0-471-85190-6
87. G.N. Afanasiev, V.G. Kartavenko and Yu.P. Stepanovsky, Vavilov-Cherenkov and transition radiations on the dielectric and metallic spheres, Journal of Mathematical Physics, 44(2003)4026–4056
88. V.L. Ginzburg and V.N. Tsytovich, On the Derivation of the Transition Radiation Intensity, Physics Letters 79A(1980)16–18
V.L. Ginzburg, On the radiation from uniformly moving sources, Nucl. Instrum. Methods Phys. Res., A: 248 (1986) 13–16
89. Charles Kittel, Introduction to solid state physics. - 7th ed. / New York, NY: Wiley, 1996
90. B. Dolgoshein, Transition radiation detectors, Nucl. Instrum. Methods Phys. Res., A: 326 (1993) 434–469
91. L.C.L. Yuan, C.L. Wang and H. Uto, Formation-Zone effects in transition radiation due to ultrarelativistic particles, Phys. Rev. Lett. 25 (1970) 1513–1515
92. J.E. Lilienfeld, Die sichtbare Strahlung des Brennecks von Röntgenröhren, Physik Zeitschrift, XX(12) 280, 1919
Mario Rabinowitz, Lilienfeld Transition Radiation Brought to Light, physics/0307047 (July 2003)
93. H. Boersch, C. Radeloff and G. Sauerbrey, Experimental Detection of Transition Radiation, Phys. Rev. Lett. 7(1961)52–54
94. X. Artru et al., Experimental investigations on geometrical resolution of optical transition radiation (OTR), Nucl. Instrum. Methods Phys. Res., A: 410 (1998) 148–158
95. J. Ruẑiẑka and J. Mehes, Properties of optical transition radiation for charged particle inclined flight through a finite thick plate IX, Nucl. Instrum. Methods Phys. Res., A: 250 (1986) 491–502, and references therein.
96. S.D. Borovkov et al., On studying the possibility to use optical transition radiation for proton beam diagnostics, Nucl. Instrum. Methods Phys. Res., A: 294 (1989) 101–107
97. J. Ruzicka et al., On optical transition radiation of charged particles in SiO_2-aerogels, Nucl. Instrum. Methods Phys. Res., A: 384 (1997) 387–402
98. X. Artru, G.B. Yodh and G. Mennessier, Practical theory of the multilayered transition radiation detector, Phys. Rev. D 16(1975)1289–1306
99. L. Fayard, Transition radiation, in: Instrumentation en Physique Nucléaire et en Pysique des Particules, les éditions de physiques, 1988, 327–340
100. J. Apostolakis et al., Parameterization models for X-ray transition radiation in the GEANT4 package, Comput. Phys. Commun. 132, 3 (2000) 241–50
V.M. Grichine and S.S. Sadilov, GEANT4 X-ray transition radiation package, Nucl. Instrum. Methods Phys. Res., A: 253 (2006) 299–302
101. F. Sauli, Principles of Operation of Multiwire Proportional and Drift Chambers, CERN 77–09
102. P. Cwetanski, Straw performance studies and quality assurance for the ATLAS transition radiation tracker, Ph.D. Thesis, Helsinki 2006, ISBN 952-10-2122-5
103. The ALICE Collaboration, K. Aamodt et al., The ALICE experiment at the CERN LHC, 2008 JINST 3 S08002
104. The ATLAS Collaboration, G. Aad et al., The ATLAS Experiment at the CERN Large Hadron Collider, 2008 JINST 3 S08003
105. The ATLAS TRT collaboration, E. Abat et al., The ATLAS TRT Barrel Detector, 2008 JINST 3 P02014

106. T. Cornelissen and W. Liebig, ATLAS Inner Detector Results from the 2004 Combined Test Beam Data, Nuclear Physics B (Proc. Suppl.) 172 (2007) 292–295
Tuan Vu Anh, Private communication, February 2009
107. ALICE collaboration, ALICE transition-radiation detector: Technical Design Report, CERN-LHCC-2001-021
108. C. Adler et al., Electron/pion identification with ALICE TRD prototypes using a neural network algorithm, Nucl. Instr. and Meth. Phys. Res. A 552 (2005) 364–371
109. M. Holder and H. Suhr, Separation of electrons and pions by a transition radiation detector using flash ADCs, Nucl. Instr. and Meth. Phys. Res. A 263 (1988) 319
110. A. Andronic, Electron identification performance with ALICE TRD prototypes, Nucl. Instr. and Meth. Phys. Res. A 522 (2004) 40–44
111. ALICE Collaboration, S. Acharya et al., The ALICE Transition Radiation Detector: construction, operation, and performance, arXiv:1709.02743v1 [physics.ins-det] 8 Sep 2017, CERN-EP-2017-222 29 August 2017

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 8
Neutrino Detectors

Leslie Camilleri

After a brief introduction describing the many sources of neutrinos, this article will describe the various detector techniques that are being used to observe neutrinos of energies ranging from a few MeV to hundred's of GeV.

8.1 Historical Introduction

In 1930 in order to explain the continuous energy spectrum of electrons emitted in beta decay, Pauli postulated [1] that these electrons were emitted together with a light neutral particle. This particle was subsequently named the neutrino. Their actual observation had to wait until 1953 when Reines and Cowan recorded [2] interactions of anti(electron)neutrinos emitted by a reactor in a cadmium doped liquid scintillator detector. Since then, in addition to the ν_e, two other flavours of neutrinos were observed, the ν_μ and ν_τ. The ν_μ, which is produced in $\pi \rightarrow \mu$ decay, was proved to be different [3] from the ν_e in an experiment at Brookhaven using thick-plate optical spark chambers. The ν_τ, companion of the τ lepton, was observed [4] at Fermilab in an emulsion cloud chamber detector consisting of iron plates interleaved with sheets of photographic emulsions. Although until recently neutrinos were thought to be massless and were described as such in the Standard Model, in the past decade they have been found to be massive [5, 6]. Furthermore each of the three flavour states mentioned above consists of a superposition of three mass states of unequal masses leading to oscillations of one flavour into another under the appropriate conditions. The characteristics of these oscillations depend on three mixing angles θ_{13}, θ_{12} and θ_{23} as well as on the difference of the square of the 3

L. Camilleri (✉)
Nevis Labs, Columbia University, Irvington-on-Hudson, NY, USA
e-mail: camil@nevis.columbia.edu

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_8

neutrino masses, Δm^2_{12} referred to as the solar mass difference as it is of importance in oscillations of solar neutrinos, $\Delta m^2_{13} \sim \Delta m^2_{23}$ referred to as the atmospheric mass difference as it drives oscillations of neutrinos produced in the atmosphere through the decay of mesons produced in cosmic ray interactions. The flavour of interacting neutrinos can only be determined if the interaction is via a charged current. In these, the ν_e, ν_μ and ν_τ respectively produce a negative electron, muon or τ lepton in the final state. Antineutrinos produce the corresponding positive charged lepton.

8.2 Sources of Neutrinos and Their Characteristics

Naturally occurring neutrinos and man-made neutrinos are produced through several different processes. Nature provides us with solar neutrinos emitted by the sun, atmospheric neutrinos produced by the interaction of cosmic rays in the atmosphere, cosmological neutrinos produced by a variety of deep space violent events, geological neutrinos produced by nuclear decays in the earth core as well as neutrinos produced in beta decay. Man made neutrinos are produced by nuclear reactors or by specially designed beams at accelerators or by highly radioactive sources. These processes are briefly described below.

8.2.1 Solar Neutrinos

They are emitted in nuclear reactions occurring in the sun [7]. The three main reactions are $\mathrm{p} + \mathrm{p} \rightarrow \mathrm{d} + \mathrm{e}^+ + \nu_e$, emitting a continuous spectrum of neutrinos with an end point at 0.4 MeV, $\mathrm{e} + {}^7\mathrm{Be} \rightarrow {}^7\mathrm{Li} + \nu_e$ with a monochromatic spectrum at 0.862 MeV and ${}^8\mathrm{B} \rightarrow {}^8\mathrm{Be}^* + \mathrm{e}^+ + \nu_e$ also with a continuous spectrum with an end point at 15 MeV. Their total flux on earth is $6.4 \times 10^{+10}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$.

8.2.2 Atmospheric Neutrinos

Atmospheric neutrinos [8] are produced in the decays of π and K mesons produced in the interactions of cosmic rays in the upper atmosphere. Their energy ranges over several orders of magnitude up to hundreds of GeV. They are observed either coming from above or from below and in the latter case they will have traversed the earth. This allows us to observe them from a few kilometers to about 13,000 km from their production point, thus providing us with very different baselines over which to study oscillations. These predominantly ν_e and ν_μ neutrinos are usually observed through their charged current interactions respectively producing electrons or muons.

8.2.3 Cosmological Neutrinos

The study of cosmological neutrinos [9] at the TeV scale is in its infancy. Their very low rate necessitates extremely large detectors. This has led to the use of naturally occurring detection media such as lake or sea water and Antarctic ice. The Cerenkov light or radio waves emitted by charged particles produced in their interactions in the medium are recorded, respectively, in strings of photomultiplier tubes or antennas.

8.2.4 Reactor Neutrinos

Nuclear reactors are an abundant source of antineutrinos, 6 $\bar{\nu}_e$ per nuclear fission on average, resulting in a flux of 1.8×10^{20} per GW thermal energy, emitted isotropically. The standard method to study them [10] is to observe the Inverse Beta Decay (IBD) reaction $\bar{\nu}_e + p \rightarrow e^+ + n$ in a hydrogen-rich liquid scintillator detector. In addition to observing photons emitted as a result of the positron annihilation, the neutron can be detected by recording photons emitted by the neutron capture in the scintillator.

8.2.5 Accelerator Neutrinos

Accelerator neutrinos are produced [11] by the decay of π and K mesons themselves produced by the interaction of a proton beam on a target as illustrated in Fig. 8.1. The target must be thick enough along the beam to maximize the proton interaction probability and yet thin enough to minimize the reinteraction probability and multiple scattering of the produced mesons such as to produce as high an energy and as focussed a beam as possible. The usual target geometry consists of a series of thin rods of low Z material such as carbon or beryllium separated by a few cms but in line with the proton beam. The mesons are then focussed by a system of toroidal magnets. These, referred to as horns [12], consist of two concentric current sheets, parabolically shaped that provide a toroidal magnetic field. Its strength is inversally proportional to the radial displacement from the beam axis and the integral is such as to bend more the particles that are further away from the beam thus providing a near parallel beam. A second horn is usually provided such as to compensate for

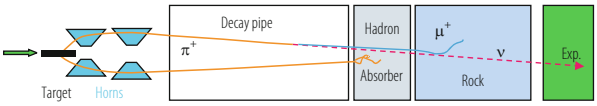


Fig. 8.1 The principle of an accelerator produced neutrino beam

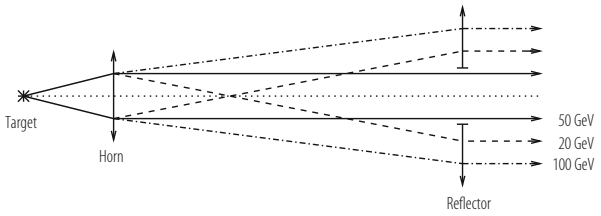


Fig. 8.2 The principle of horn focusing of mesons in a neutrino beam

over-focussed and under-focussed particles as shown in Fig. 8.2. The particles then enter a long evacuated decay tunnel in which $\pi \rightarrow \mu\nu_\mu$, $K \rightarrow \mu\nu_\mu$ and $K \rightarrow \pi e\nu_e$ decays occur producing a predominantly ν_μ beam with an admixture of $\sim$1% of ν_e. Focussing positive mesons produces a neutrino beam whereas focussing negative mesons, achieved by a reversal of the polarity of the horns, produces an antineutrino beam.

An alternative to the horns is a system of bending magnets and quadrupoles. Such a technique [13] has been used in the Sign Selected Quadrupole Train, SSQT, neutrino beam at Fermilab. Its performance is described in [14]. An advantage of this technique is that the neutrino beam not being along the axis of the proton beam, ν_e from K°_L decays will not enter the detector since their parents will not be deflected. This is a distinct advantage in oscillation experiments looking for ν_e appearance in a ν_μ beam in which the intrinsic ν_e background is irreducible.

The above techniques produce a beam with a broad energy spectrum, referred to as a broad band beam. A narrower range of neutrino energies is sometimes desirable. Such narrow band beams are obtained by first momentum-selecting the parent pions and kaons before they decay using standard beam optics methods, thus reducing the range of neutrino energies. Furthermore, the neutrino energy can be deduced on an event by event basis as it is related to the neutrino production angle and this can be computed from the radial position of the event within the detector. The uncertainty on the energy depends on the momentum and angular spread of the meson beam and on the length of the decay channel. It is typically 5–20%. The intensity of these narrow band beams is necessarily lower than that of broad band beams.

Another way to expose the detector to neutrinos with a given narrow energy spectrum is to place the detector at an off-axis angle to the beam [15]. The kinematics of pion decay, shown in Fig. 8.3 are such that neutrinos observed at a non-zero angle to the proton beam have an approximately unique momentum irrespective of the momentum of their parent meson. Furthermore the value of this unique momentum depends on the off-axis angle, thus allowing a detector to be exposed to the neutrino momentum required by the physics under investigation by placing it at the appropriate angle.

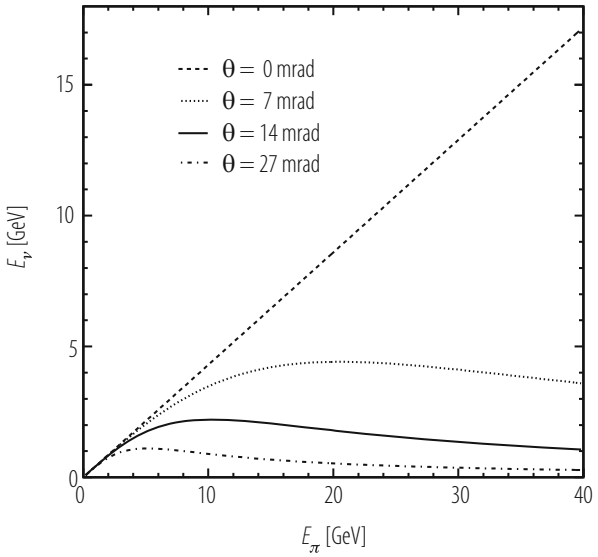


Fig. 8.3 The correlation between the pion momentum and its decay neutrino momentum, plotted for several neutrino directions relative to the proton beam axis

An accelerator neutrino beam can also, potentially, contain ν_τ's. These are produced through the production of D_s mesons in the initial proton interactions and their subsequent decays $D_s \rightarrow \tau\nu_\tau$ followed by $\tau \rightarrow \nu_\tau + \ldots$. However in most accelerator beams the ν_τ content is negligible since the D_s production cross section is small at existing energies. One notable exception will be discussed in Sect. 8.3.4.

The semi-leptonic decays of charmed particles have also been used to produce neutrinos. In this case, because of the very short lifetime of charm, a decay tunnel is unnecessary. The beam is produced in a so-called beam dump [16], in which the incident proton beam and secondary pions and kaons are absorbed before they can decay. At CERN, the beam dump [17] was made of copper disks that could be separated thus altering its density between 3 and $9\,\mathrm{g}\cdot\mathrm{cm}^{-3}$. The normal neutrino flux from π and K decays was reduced by about 3 orders of magnitude. Since charm is produced in pairs in proton interactions, the beam contains an approximately equal amount of neutrinos and antineutrinos, the only difference being due to π and K mesons decays occurring before these mesons are absorbed. Furthermore, because of the equal $e\nu_e$ and $\mu\nu_\mu$ decay probabilties of charm an equal number of ν_e and ν_μ are present in the beam. In this context, it is interesting to note [18] that hadronic colliders, and in particular LHC, produce a large quantity of charm and beauty particles in the forward directions, resulting in two well collimated neutrino beams emerging from each interaction region.

8.3 Detection Techniques

Because of the small interaction cross section of neutrinos, neutrino detectors must be massive. The exception is detectors addressing coherent neutrino interactions for which the cross section is orders of magnitude larger than for other neutrino interactions and which will be addressed in Sect. 8.3.1. The nature of these massive detectors depends on the physics being addressed. It usually involves observing the resulting hadronic part of an interaction and, if a charged current interaction, the observation of a charged lepton. If the physics merely requires the measurement of the total neutrino energy, a calorimetric detector suffices. If individual particles must be measured, then a more sophisticated tracking device is needed. The measurement of a final state muon is usually accomplished in a fairly straightforward way with magnetized iron because of the muon penetrating nature. A final state electron is more difficult to measure, especially its charge, because of bremsstrahlung and showering as it propagates through material. Neutrino detectors must then necessarily be of several types. Techniques must be suitable to detect neutrinos of energies ranging from a few MeV to about a PeV. They must be fine-grained enough to measure electrons, identify individual particles and observe secondary vertices of τ's or charmed particles or heavy enough to produce large number of interactions using calorimetric techniques. It is evident that neutrino detectors use most of the detecting techniques used in particle physics. They will be outlined in the following sections.

8.3.1 Totally Active Scintillator Detectors

Scintillator detectors can either use liquid or solid scintillator. If liquid is used the detector consists of either a single large tank or of tubes filled with liquid. Solid scintillator detectors usually consist of strips. The first neutrino detector [2] used by Reines and Cowan was intended to observe the interaction of reactor antineutrinos of a few MeV. The observation was made, as in subsequent reactor experiments, using the IBD reaction $\bar{\nu}_e + p \rightarrow e^+ + n$ and using a detector consisting of liquid scintillator viewed by photomultipliers. In addition to observing light emitted from the positron annihilation, the neutron can be detected by also observing photons emitted by the neutron capture in the hydrogen of the scintillator. In order to enhance the neutron capture cross section, they added cadmium to the scintillator. They observed an excess of events when the reactor was in operation leading to the first detection of (anti)neutrino interactions and a subsequent Nobel prize. This technique is still being applied to this day [10], albeit with some refinements. Several recent experiments which will be described below, used it to search for $\bar{\nu}_e$ oscillations to another flavour in the domain of the atmospheric $\Delta m^2, 2.5 \times 10^{-3}\,\mathrm{eV}^2$. Because of the low energy of reactor $\bar{\nu}_e$'s, $\bar{\nu}_\mu$'s or $\bar{\nu}_\tau$'s that they potentially oscillate to cannot be observed through their charged current

interactions since it is energetically impossible to produce μ's or τ's. Oscillations can then only be observed through the disappearance technique resulting in a reduction and distortion of the expected $\bar{\nu}_e$ spectrum. Given the energy of reactor $\bar{\nu}_e$'s (a few MeV) and the value of the atmospheric Δm^2, CHOOZ [19] was located 1000 m from a reactor complex in order to be near oscillation maximum. It used a single large tank of liquid scintillator and was subjected to a cosmic muon rate of $0.4\,\mathrm{m}^{-2}\mathrm{s}^{-1}$. One of the major backgrounds in this type of experiment is the background generated by cosmic ray muons. The first line of defense is to place the detector underground, at a depth of 300 m water equivalent (m.w.e) in the case of CHOOZ. A muon traversing the detector does not, in itself, simulate a signal event because the large amount of energy deposited can be well identified. However neutrons produced by muons traversing dead areas of the detector or the surrounding rock can elastically scatter on a proton, causing the proton and the subsequent neutron capture to simulate the signature of a reactor event. This background can be eliminated by vetoing on the passage of a nearby muon. In addition cosmic muons can produce long lived isotopes such as ^{6}He and ^{9}Li which subsequently can beta decay producing an electron and a neutron, thus simulating an antineutrino event. This background cannot be eliminated by vetoing on the passage of a muon because of the long lifetime of these decays (178 ms in the case of ^{9}Li) which would introduce an inordinate dead time. It must be estimated and subtracted. Palo Verde [20] was located at a shallower depth of 32 m.w.e. and chose to use acrylic cells filled with liquid scintillator. This extra segmentation was needed to reduce the larger muon induced neutron background caused by the larger cosmic muon flux of $22\,\mathrm{m}^{-2}\mathrm{s}^{-1}$ at this depth. Instead of cadmium, these experiments have been using a 0.1% admixture of gadolinium with a large neutron absorption cross section leading to an 84% capture fraction. Absorption in gadolinium leads to the emission of gamma rays with a total energy of 8 MeV, within $\sim$30 µs and $\sim$6 cm of the positron annihilation, thus providing a well recognizable delayed coincidence. The CHOOZ target scintillator consisted of 50% by volume Norpar-15 [21] and IPB + hexanol (also 50% by volume). The wave-length shifters were p-PTP and bis-MSB (1 g/l). The gadolinium was introduced as a solution of $Gd(NO_3)_3$ in hexanol. Because of oxygenation of the nitrate the 4 m light attenuation length in the scintillator decreased with time at a rate of $(4.2 \pm 0.4) \cdot 10^{-3}$ per day. This required a careful monitoring of the scintillator transparency using calibration sources. The light yield was 5300 photons/MeV.

The observation of a modification of the expected $\bar{\nu}_e$ spectrum necessitates a very precise knowledge of the antineutrino flux emitted by the reactor as well as of the antineutrino interaction cross section. They failed to observe a disappearance of antineutrinos and the limit set on this oscillation was governed by these sources of systematics uncertainty. In order to overcome these limitations more recent reactor oscillation experiments use a second identical detector located close to the reactor in order to measure the expected interaction rate before the neutrinos have a chance to oscillate. The detector used by one such experiment, Double Chooz [22], located at the same location as CHOOZ but using, in addition, a near detector placed at 410 m from the reactors, will be described as an example. The scintillator, amounting to

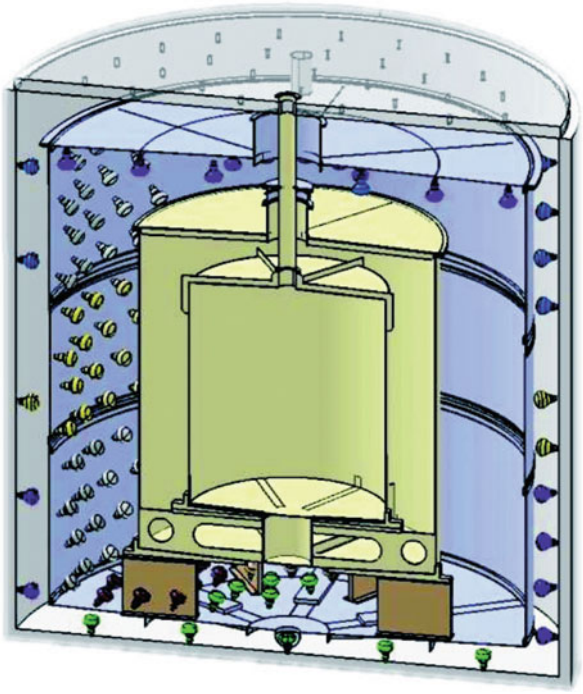


Fig. 8.4 The Double Chooz detector

10 tons, is housed in a central tank, Fig. 8.4, consisting of a clear acrylic. It is surrounded by a gamma catcher consisting of undoped liquid scintillator housed in a second acrylic shell that provides additional gamma detection probability for interactions occurring near the boundary of the central tank. A third envelope consisting of stainless steel holds the photomultipliers that view the two scintillator volumes through the acrylic shells. A buffer consisting of mineral oil fills the space between the stainless steel shell and the acrylic shell of the gamma catcher. It serves the purpose of absorbing any radioactivity emitted by the photomultipliers. In order to veto on cosmic muons the photomultiplier shell is itself surrounded by yet another scintillator volume housed in a final outer shell that also holds a second set of photomultipliers viewing this inner veto layer. Lastly, planes of scintillator counters cover the ceiling of the detector cavern, to identify more muons that traverse the surrounding material and are a potential source of neutrons. The scintillator chosen for the target is a 20/80 mixture of phenyl-xylelethane (PXE)/dodecane with 0.1% gadolinium doping introduced as a dipivaloymethane molecule, $Gd(dpm)_3$. This has demonstrated long term stability. With PPO and Bis-MSB as fluors the mixture has an attenuation length greater than 5 m at 450 nm and a light yield of 7000 photons/MeV resulting in 200 detected photoelectrons/MeV. The positron detection

threshold is less than 700 keV, well below the threshold of 1.022 MeV of the inverse beta decay reaction.

Calibration of the detector is required to determine, in both detectors, the efficiency for observing the inverse beta decay reaction, the energy scales for positrons, neutrons and gamma, the timing of the photomultipliers and the light transport properties. To this end gamma sources, neutron sources and laser light flashers are used and deployed throughout the detector volumes in order to map out the relevant parameters. In the target this is done with an articulated arm at the end of which is mounted the calibrating device, the position of which is determined by the length and azimuthal position of the arm. In the gamma catcher a guide tube into which a source can be inserted at the end of a wire is used. The geometry of the tube and the wire length determine the position of the source.

Whereas Double Chooz was the first to report a hint for a non-zero θ_{13}, two experiments have since produced the best measurements of this angle. They use the same concept as Double Chooz but have used either a larger neutrino flux (RENO [23]) or more detectors and more flux (Daya Bay [24]). RENO, in South Korea, is exposed to the flux of 6 reactors in a row totalling 16.4 GW_{th}. Its far detector is 168 m underground and 1380 m from the central reactor whereas its near detector is 46 m underground and 290 m from the reactor line. Its inner target weighs 15.4 tons and is viewed by 340 photomultipliers. Daya Bay, in China, uses eight identical detectors and is located near three reactor complexes Daya Bay, Ling Ao I and II, a total of 17.4 GW_{th}. Its far detector hall is 324 m underground, 1540 m from Ling Ao and 1910 m from Daya Bay and houses four detectors. One near detector hall 363 m from the Daya Bay complex and another one about 500 m from the Ling Ao complex each house 2 detectors. Each detector includes a 20 ton neutrino target viewed by 192 $8''$ photomultipliers. Their inner vetos are tanks of water in which Cerenkov light is viewed by photomultipliers. The Daya Bay energy resolution is $\sigma_E/E = 7.5\%/\sqrt{E}$. Using a variety of radioactive sources they are able to determine the absolute neutrino energy scale to 1% and the relative energy scale between detectors to <0.2%. The relative detection efficiency uncertainty was 0.13% and was substantiated by comparing rates of detectors in the same hall. Table 8.1 compares the systematic uncertainties achieved in Daya Bay to the ones in the CHOOZ single detector experiment demonstrating the effectiveness of a multiple detector and multiple location experiment. It should be noted that all three experiments have observed a structure in the positron energy distribution between 4 and 6 MeV when compared to Monte Carlo predictions based on the present understanding of a reactor neutrino flux. This structure is also seen in their near detectors (see for instance [25]) and its amplitude is proportional to the reactor flux. It is therefore believed to be due to our lack of complete understanding of the complex origin of a reactor neutrino flux.

KamLAND [26] is also an experiment observing reactor antineutrinos but studies oscillations in the domain of the solar $\Delta m^2, 7.5 \times 10^{-5}\,\mathrm{eV}^2$, and is therefore situated at an average distance of about 180 km from 53 Japanese power reactors to compensate for the much smaller Δm^2. The same reaction and technique as described above are used. However to observe enough events at this distance

Table 8.1 The Daya Bay systematic uncertainties compared to the ones in CHOOZ

Variable	CHOOZ [%]	Daya Bay [%]
ν flux and cross section	1.9	–
Reactor power	0.7	–
Energy per fission	0.6	–
Number of protons	0.3	0.03
H/C ratio and Gd concentration	1.2	0.1
Spatial effects	1.0	0.02
Live time	–	0.01
Analysis cuts	1.5	0.082
Total	2.7	<0.13

a detector consisting of 1 kiloton of liquid scintillator had to be used. The scintillator is housed in a 13 m diameter transparent nylon balloon suspended in non-scintillating oil acting as a buffer and viewed by 1879 photomultipliers. This inner detector is surrounded by a 3.2 kiloton water Cerenkov counter which has the dual purpose of reducing γ rays and neutrons from the surrounding rock and of detecting cosmic ray muons. As well as measuring the oscillation pattern as a function of L/E of reactor neutrinos within their detector, KamLAND also made a measurement of geological neutrinos [27].

The KamLAND detector would also be used in the IsoDAR project [28] searching for sterile neutrinos through $\bar{\nu}_e$ disappearance at a Δm^2 of $\sim$1 eV^2. A 60 MeV cyclotron would be placed a few meters from the surface of KamLAND and 16.5 m from its centre, Fig. 8.5. The cyclotron would accelerate H_2^+ ions (a hydrogen molecule with one electron removed) as the single charge for two protons of H_2^+ reduces the repulsive force within a bunch and hence minimizes the effect of space charge blow up of the beam which in turn keeps beam loss down. A high current source, currently under commissioning, would produce the H_2^+ ions which would be bunched with a radio-frequency quadrupole placed vertically above the centre of the cyclotron. The bunched ions would be bent electrostatically into the plane of the cyclotron. After 96 turns they would be extracted with a thin septum, stripped and transported 50 m, resulting in 10 mA of protons impinging on a beryllium target placed near the KamLAND detector. Neutrons produced in this target stream through a sleeve consisting of small beryllium spheres surrounded by highly enriched (99.995%) ^{7}Li. A graphite reflector surrounds the target and sleeve. The neutrons captured by the lithium produce ^{8}Li which subsequently decays producing $\bar{\nu}_e$'s entering the KamLAND detector in which they can be detected via IBD. The (12 cm/$\sqrt{E_{MeV}}$) spatial resolution and (6.4%/$\sqrt{E_{MeV}}$) energy resolution of KamLAND coupled with the detector size allows the observation of event rate oscillations as a function of L/E within KamLAND in addition to an overall disappearance of $\bar{\nu}_e$'s.

Reactor complexes are next planned to be used as sources of antineutrinos to illuminate two larger versions ($\sim$10 kilotons up from $\sim$10 tons) of the current

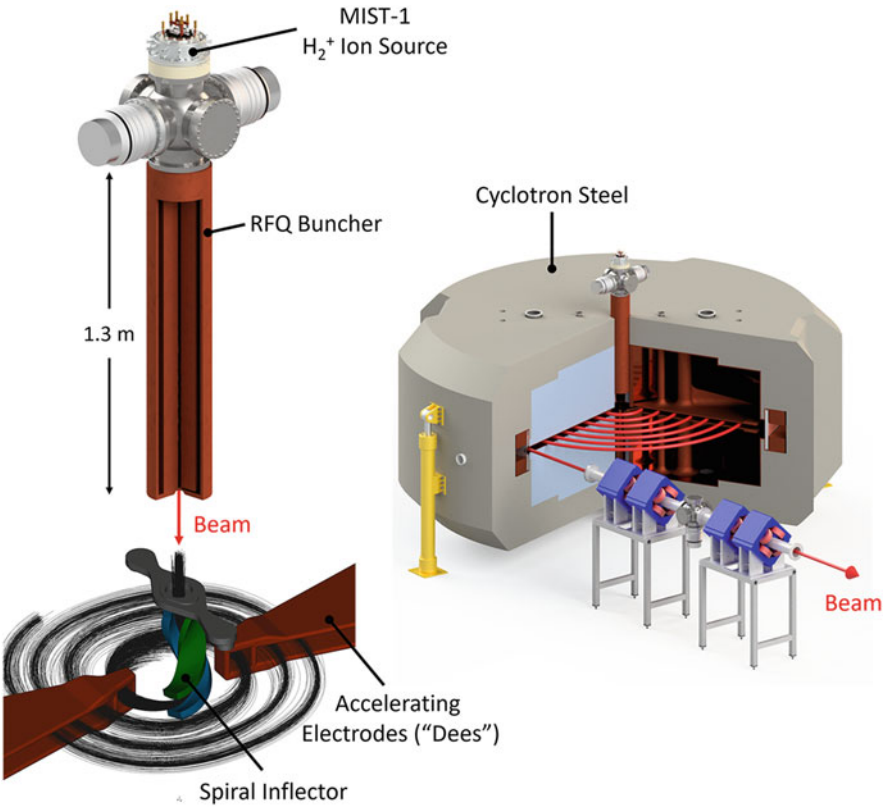


Fig. 8.5 The layout of the IsoDAR 60 MeV cyclotron and its input stage: H_2^+ source, RFQ and inflector. The same layout is to be used as an input stage for the DAEδALUS project (see Sect. 8.4.3)

reactor detectors. These detectors RENO50 [29] and JUNO [32] would be located ~50 km from the reactors. At this distance the disappearance of reactor $\bar{\nu}_e$'s is dominated by the solar Δm^2. However the additional small effect arising from the atmospheric Δm^2 is mass hierarchy dependent. A precise energy measurement of the IBD positron therefore allows the determination of the mass hierarchy, as well as more accurate measurements of θ_{12} and Δm^2_{12}.

The possibility to deploy a 10 kiloton liquid scintillator immersed off shore in the vicinity of a nuclear reactor complex in order to perform oscillation physics has been investigated [33]. This Hawai Anti-Neutrino Observatory, Hanohano, could alternatively be deployed far from a reactor in order to observe geological neutrinos.

Several scintillator detectors are also planned for deployment very close to reactors for neutrino oscillation into a sterile neutrino, accurate flux measurements and reactor monitoring, all using the IBD reaction. For this purpose they need to be compact. They also need to be segmented to mitigate the background from reactor

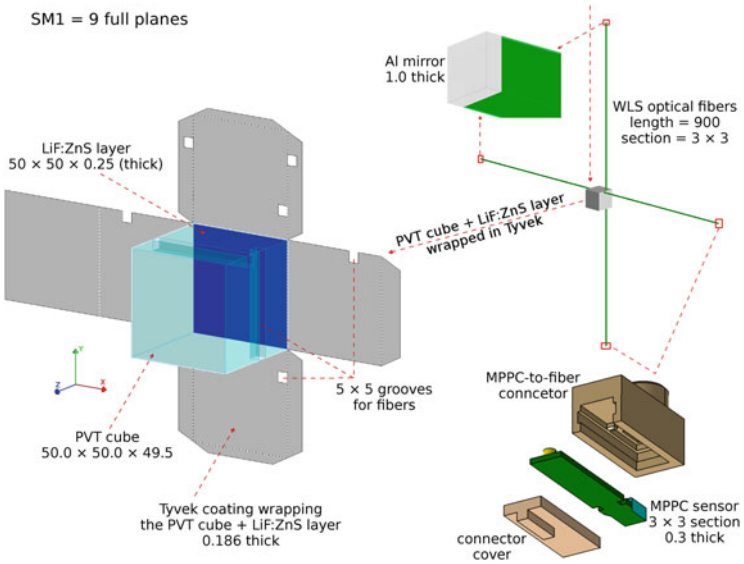


Fig. 8.6 An exploded view of a single PVT cube used in the SoLid detector prototype

neutrons. STEREO will use the same detector technique as described above for the θ_{13} experiments. Prospect [30] will run in 2 phases near a reactor at ORNL. Phase I will use a 3 ton single volume ^{6}Li loaded liquid scintillator detector movable between 7 and 12 m. The scintillator is EJ-309 from Eljen Technology to which ^{6}Li, PPO fluors and bis-MSB wavelength shifters have been added resulting in 6500 detected photons/MeV and a 4 m attenuation length. The volume is segmented by low mass optical separators into 120 segments 14.4 × 14.4 cm in cross-sectional area and 120 cm long, read by a pmt at each end. The positron deposits its energy in the liquid scintillator and the neutron is observed via its capture in hydrogen or ^{6}Li. The phase II detector will have a larger 10 ton mass while maintaining the same segmentation geometry and cover baselines between 15 and 19 m. Another example is the 1.6 ton SoLid built after prototyping a 288 kg version [31] deployed near the Belgian BR2 reactor. The final detector is built out of 12,000 $5 \times 5 \times 5\,\text{cm}^3$ ELJEN Technology EJ-200 polyvinyl toluene (PVT) cubes, Fig. 8.6. Sheets of ^{6}Li:ZnS(Ag), 225 µm thick allow the detection of the IBD neutrons through break up of the lithium to an alpha and ^{3}H with a Q-value of 4.78 MeV. The signals of each cube are read through two wave length shifting fibres connected to Hamamatsu S12572-050P multi-pixel photon counters.

Borexino [34] is a detector installed in Italy at the Laboratorio Nazionale del Gran Sasso (LNGS) for the measurement of solar neutrinos and in particular the ^{7}Be 862 keV monochromatic line using the purely leptonic reaction $\nu + e \rightarrow \nu + e$. This reaction results in an electron spectrum with a sharp edge at 665 keV. Borexino consists of 300 tons of liquid organic scintillator (pseudocumene and $1.5\,\text{g} \cdot \text{l}^{-1}$ PPO as fluor) housed in a nylon vessel itself suspended in a stainless

steel sphere on which are mounted 2200 photomultipliers. The sphere is filled with a pseudocumene solvent with a quencher acting as a shield for radioactivity coming mainly from the tubes and is itself immersed in a water Cerenkov tank viewed by an additional 200 photomultipliers to identify cosmic ray muons. The light yield of the detector is 500 photoelectrons/MeV actually recorded. Timing information from the photomultipliers allow the spatial reconstruction of the event and hence the determination that it occurred within the fiducial volume. The α and β^+ components of natural radioactivity can be reduced by pulse shape discrimination whereas the β^- and γ components are indistinguishable from the signal. A reduction and thorough understanding of the background has allowed them to observe solar neutrinos with an energy as low as 150 keV and, hence, make the first direct observation of pp fusion solar neutrinos as well as measure the solar beryllium line and geoneutrinos [35]. After a year during which the background has been reduced through six cycles of water extraction the radiopurity levels are now 2.7×10^{-18} for $^{14}C/^{12}C$ and, at 95% CL, $<9.7 \times 10^{-19} g \cdot g^{-1}$ for uranium and $<1.2 \times 10^{-18} g \cdot g^{-1}$ for thorium. This will allow improved measurements of solar and geoneutrinos as well as a new very short baseline neutrino oscillation project, SOX [36].

SOX is intended to search for oscillations of decay ν_e or $\bar{\nu}_e$ from a radioactive source into sterile neutrinos at the level of Δm^2 of 1 eV^2. The sources being considered are ^{51}Cr and ^{144}Ce, with the latter already approved. The ^{144}Ce would be placed in a pit under the Borexino detector. The small size of the 3.7–5.0 PBq source (about 1 L) coupled with the large 7 m size of the detector and its good spatial resolution of 12 cm and energy resolution of 3.5% would allow the observation of oscillation waves as a function of L/E within the detector as well as an overall measurement of $\bar{\nu}_e$ disappearance. Given an existing detector, the most taxing task is the source. It would be produced in a Russian laboratory from the reprocessing of nuclear fuel and must then be extensively shielded and transported to the Gran Sasso by a circuitous route for safety reasons.

Totally active liquid scintillator detectors have also been used in accelerator experiments producing higher energy neutrinos. MiniBooNE [37], looking for $\nu_\mu \rightarrow \nu_e$ oscillations in the Fermilab Booster neutrino beam in order to investigate the LSND signal [38], is exposed to neutrinos of about 1 GeV. The detector consists of 800 tons of mineral oil (CH_2) held in a spherical tank. The density of the oil is $0.86\ g \cdot cm^{-3}$ and has an index of refraction of 1.47. The light attenuation in this medium varies from a few cm at 280 nm to 20 m at 400 nm. The inner region (575 cm radius) is viewed by 1280 8-inch photomultipliers held on an optical barrier that separates it from a 35 cm thick outer region. This outer region, itself viewed by 240 tubes is used to veto events caused by charged particles entering the detector and to tag events that include particles exitting the detector in order to identify contained events. Cosmic ray events are greatly reduced by restricting the triggers to those occurring within a 19.2 µs window starting 4.4 µs before the 1.6 µs long beam spill. The energy of an event is related to the total amount of light observed. ν_μ and ν_e events are identified by the flavour (muon or electron) of the lepton in the final state CC interaction. Muons are distinguished from electrons using the light pattern of their Cerenkov rings as shown in Figs. 8.7 and 8.8. Muons give a sharp ring filled on

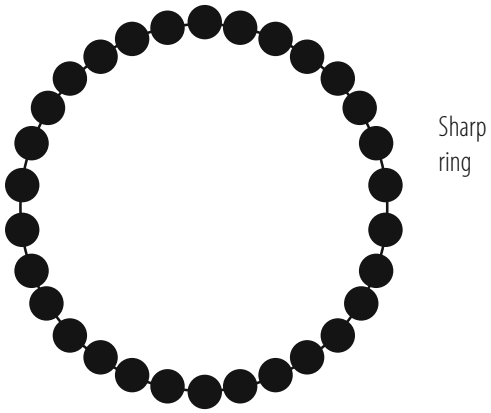


Fig. 8.7 The Cerenkov light pattern characteristic of a muon in the MiniBooNE detector

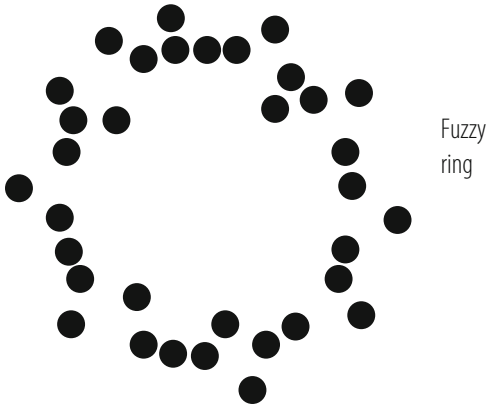


Fig. 8.8 The Cerenkov light pattern characteristic of an electron in the MiniBooNE detector

its interior as the muon approaches the tubes. Electrons give a fuzzy ring because of the many electrons and positrons each moving in a slightly different direction within the showers. In addition to the intrinsic ν_e component of the beam caused by μ, K and π decays the background to the ν_e appearance search comes from π^0 decays to two photons. This background can be greatly reduced by the ability of the detectors to observe separately the two electron-like rings produced by the two photons. The π^0's can be reconstructed with a mass resolution of 20 MeV/c^2. The event vertex, direction and energy resolutions with which ν_e events are reconstructed are 22 cm, 2.8° and 11% respectively. The experiment observed an unexplained excess of electromagnetic low energy events, but was not able to determine whether they were due to single photons or electrons due to the similarity of the rings produced by them.

Liquid scintillator detectors can also be of a tracking kind, in which the scintillator is confined in tubes and read by wave length shifting (WLS) fibres. NOvA [39], an experiment that runs in the Fermilab 2 GeV off-axis NuMI beam at a distance of 810 km from the lab is such an example. It consists of planes of extruded PVC tubes alternating in the horizontal and vertical direction. Each tube is 3.87 cm by 6 cm in cross-sectional area, 15.6 m long and is filled with mineral

oil with 5% pseudocumene. Since one of its physics goals is ν_e appearance it must be fine-grained enough to identify electrons and distinguish them from the showers produced from the decay photons of π^0 mesons. To this end, each plane corresponds to a sampling frequency of only 17% of a radiation length. The WLS fibres are in the shape of a loop and are read at the end opposite the loop. Avalanche photodiodes with a quantum efficicency of about 80% are used and detect 40 photoelectrons for a minimum ionizing particle crossing a tube at the far end. They are produced in arrays of 16 diodes each with a cross-section of $1.8 \times 1.0\,\text{mm}^2$ and must be run at a temperature of $-15\,^\circ\text{C}$. Each diode observes both ends of a fibre. The detector has an overall mass of 14 kilotons, consisting 70% of scintillator and the remainder of PVC. Its overall length is 67 m, with a cross section of $15.7 \times 15.7\,\text{m}^2$. The detector is located on the surface but the impact of cosmic rays is mitigated by the short beam spill of 10 µs and the speed of the photodiodes. Nonetheless, to reduce the electromagnetic component of cosmic rays, the detector is covered by a 3 m overburden of concrete and barite.

Totally active tracking detectors can also be made of extruded solid scintillator bars usually read with WLS fibres embedded in a hole or a groove made in the scintillator. An example of such a detector is SciBar [40], a 15 ton detector consisting of 14,336 strips each of dimensions $1.5 \times 2.5 \times 300\,\text{cm}^3$ and using 64-pixel multianode photomultipliers. It was first used in Japan on the KEK neutrino beam line and then moved to the NuMI beam line at Fermilab in the US.

Coherent Elastic Neutrino-Nucleus Scattering, CEνNS , is a process in which the neutrino interacts with the whole nucleus rather than with individual nucleons [41], leaving the nucleus whole and carrying very little energy since the momentum transfer must be small. The recoiling nucleus subsequently produces secondary recoils and scintillation light. The CEνNS cross section is several orders of magnitude larger than, for instance, the IBD cross section as it depends on the square of the number of neutrons in the nucleus. However, the smallness of the energy release made the process impossible to measure until recently. The COHERENT experiment [42] overcame this difficulty by using a 14.6 kg sodium-doped CsI crystal 34 cm long. The heavy cesium and iodine nuclei, provide the large cross sections and the large scintillation light yield necessary to detect low energy recoil nuclei down to a few keV. The crystal was read by a super bialkali low background Hamamatsu R877-100. The source of neutrinos was the decay of pions and muons produced by the Spallation Neutron Source at Oak Ridge National Laboratory. Protons on target(POTs) were delivered in 1 µs long spills at a rate of 60 Hz resulting in 4×10^{18} isotropically emitted neutrinos per day. The detector was placed in a basement corridor at a location, Fig. 8.9, that provided 12 m of neutron-moderating concrete and gravel in the direct line of sight to the SNS target, thus reducing neutron-induced recoil nuclei background (NIN) to an acceptable level. Cosmic rays were also reduced with an 8 m water overburden. The detector was enclosed in high-density polyethylene, to reduce NIN, as well as in both low activity and in standard lead. Muon vetos and water tanks containing a neutron moderator completed the shielding of the detector. The photomutiplier signals were amplified and digitized at 500 MSamples/s over 70 µs intervals starting 55 µs before the POT

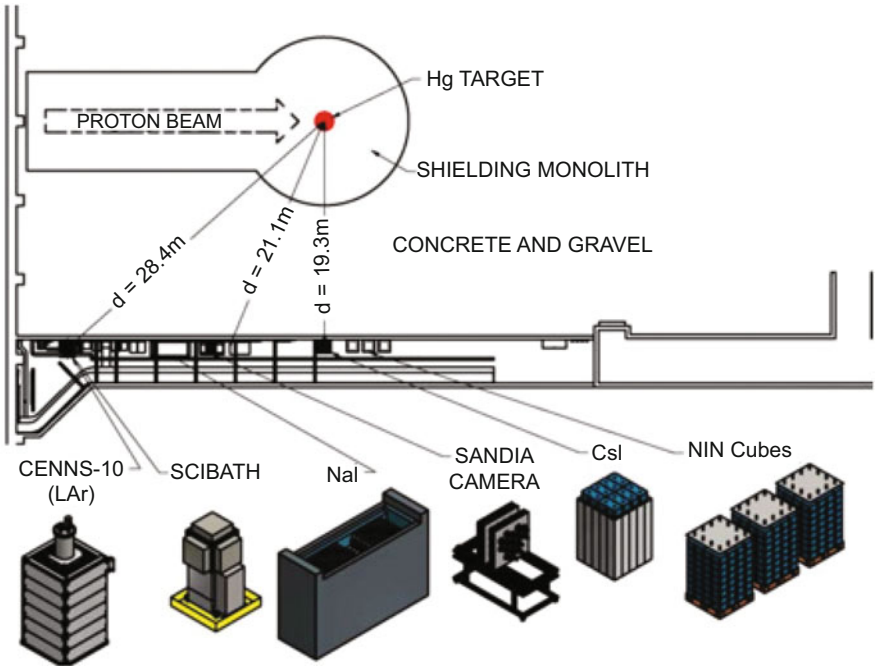


Fig. 8.9 The COHERENT experiment layout, showing the proton beam, target, shielding and experimental area

signal. Two 12 µs windows, one preceding and one following the POT trigger, allowed the comparison of data respectively unrelated and related to the beam. The 40 µs interval preceding these two windows was used to veto events due to previous energy depositions. The window following the POT signal showed a distribution of events consistent both in energy with coherent scattering and in time distribution with pion and muon decays. These signals were absent in the window preceding the POT, allowing the experimenters to announce a first observation of CEνNS at the 6.7 standard deviation confidence level. Detectors with different technologies such as liquid argon and NaI[T1] crystals are currently installed in the same location with further expansions being considered.

8.3.2 Water Cerenkovs

These are large volumes of ultra-pure water in which charged particles produced in neutrino interactions are detected through the Cerenkov light they emit. The measurement and separation of electrons, muons and π^0's is as was described in the context of MiniBooNE and illustrated in Figs. 8.7 and 8.8. Several experiments [43, 44], originally conceived to search for proton decay have pioneered this technique

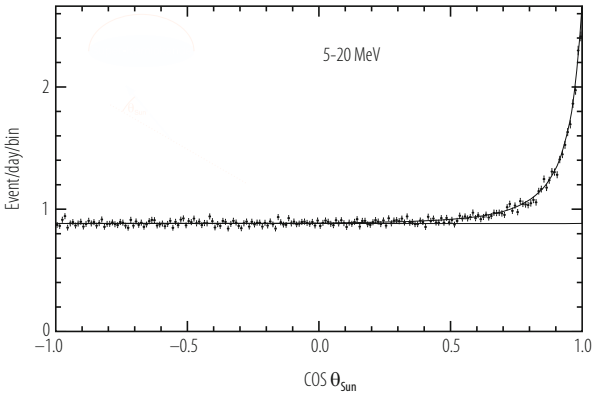


Fig. 8.10 The electron direction relative to the sun position in solar neutrino candidate events observed in SuperKamiokande

for the study of atmospheric and solar neutrinos, and, as a byproduct, have made the first recording of neutrinos emitted by a supernova, namely SN1987A. These detectors have been followed by the most productive one, SuperKamiokande(SK) [45], a 50 kiloton detector placed in the Kamioka mine in Japan at a depth of 2700 m water equivalent. It consists of two concentric cylindrical detectors. The inner one (ID) is viewed by 11,146 photomultipliers of 20 inch diameter providing a 40% coverage while the outer one(OD) is viewed by 1885 8 inch tubes. The absence of signal in the OD distinguishes fully contained events from partially contained ones. SK has successfully observed neutrinos ranging in energy from a few MeV(solar neutrinos) to several tens of GeV (atmospheric and accelerator neutrinos). The electron energy and direction produced in the elastic scattering reaction used to observe solar neutrinos are related to the incident neutrino energy and direction. The pointing accuracy is such that the origin of these neutrinos can be clearly associated to the sun, Fig. 8.10. In order to observe as much of the solar neutrino spectrum as possible, it has minimized background such as to be able to lower their detection threshold to ∼3.5 MeV. They have observed neutrino interactions coming from above and from below and have observed the reduction of ν_μ interactions from below (long baseline) due to $\nu_\mu \rightarrow \nu_\tau$ oscillations. Using a neural network approach they were also able to identify the resulting ν_τ CC component. SK is currently the heart of the T2K long baseline experiment [46] in which it is exposed to a beam of neutrinos or antineutrinos from the JPARC accelerator laboratory 295 km away. The beam is produced starting with 30 GeV protons and is an off-axis beam with a narrow neutrino energy spectrum peaked at 600 MeV. The experiment includes a near detector which will be described in Sect. 8.3.5. Their excellent electron and muon identification have allowed them to observe ν_μ disappearance as well as ν_e appearance, and measurements of $\sin^2\theta_{23}$, Δm^2_{23} as well as θ_{13}. In long baseline ν_e appearance experiments such as T2K or NOvA, the measured value of θ_{13} is correlated to the yet unknown CP violation phase, leading these experiments to

quote their measurement of θ_{13} as a range of values driven by the allowed CP phase values. Combining this range of θ_{13} with the precise reactor experiments measurement of θ_{13}, allows to limit the possible values of the CP violation phase.

Several long baseline projects based on the water Cerenkov technique have been proposed to measure the mass hierarchy and the CP phase via $\nu_\mu \rightarrow \nu_e$ oscillations. However because of the low signal event rate expected, detectors of the order of the megaton are needed. This would enable these detectors to continue the very successful non-beam physics programme of SK, namely atmospheric and solar neutrino physics, proton decay searches and supernovae watches. The MEMPHYS [47] project was planned in the context of a potential neutrino beam [48] from CERN to a new laboratory in the Frejus tunnel. It consists of 3 cylindrical water Cerenkov counters placed in contiguous caverns for a total of 0.5 megatons. Hyper-Kamiokande(HK), described in their Letter of Intent [49], is a natural extension of SK and would use the same beam as SK (600 MeV off-axis at 2.5°) but with its power upgraded from 0.75 MW to 1.35 MW, mostly by increasing the JPARC main ring repetition rate to 0.86 Hz. In their latest design the beam would impinge on a 0.52 Mton (0.38 Mton fiducial) water Cerenkov detector consisting of two 74 m diameter and 60 m high cylinders located 295 km from J-Parc in a cavern 650 m underground and 8 km south of SK. The Cerenkov light would be observed by 80,000 50 cm diameter Hamamatsu R12860 photomultipliers of a new Box and Line design providing 40% coverage with single photons detected with a 24% efficiency and a 1ns timing resolution. The photomultipliers have survived extensive pressure and implosion tests. Alternative sensors are also being investigated such as Hybrid Photo Detectors and the multi-photomultipliers concept developed for KM3Net [72]. The extrapolation of the water Cerenkov technique to a megaton-sized detector is driven to a large extent by the cost of the optical sensors and their production schedule, making the spacing and size of the sensors of prime importance. Similar detectors were considered for installation at SURF DUSEL [50] in order to observe neutrinos from a new beam from Fermilab. However, as will be discussed below, the liquid argon technique has been adopted instead.

A test [52] in the Super-Kamiokande detector demonstrated that neutrons from the IBD reaction $\bar{\nu}_e + p \rightarrow e^+ + n$ could be detected with the addition of gadolinium in the water. A 2.4 L acrylic vessel was filled with a 0.2% $GdCl_3$ mixture. A BGO crystal containing an Am/Be radioactive source was placed in its middle. The α particles emitted by the ameritium interacted in the beryllium via $\alpha + {}^9\text{Be} \rightarrow {}^{12}C^* + n$. By immersing the vessel in the SuperKamiokande detector and using the 4.43 MeV carbon deexcitation photon as a trigger it was demonstrated that the neutron could be detected via its absorption in the gadolinium, as described earlier in Sect. 8.3.1, with an efficiency of 66.7% with a 3 MeV threshold for delayed events. It was estimated that a background reduction of 2×10^{-4} could be achieved at a 10 MeV prompt event analysis threshold for $\bar{\nu}_e$. This opens the way for the use of the water Cerenkov technique for the detection of $\bar{\nu}_e$'s of geological or reactor origin.

Detectors that measured a deficit in the solar neutrino spectrum were all sensitive to ν_e only. In order to definitely prove that the deficit was due to a flavour

transformation rather than a disappearance it remained to prove that the overall flux of neutrinos, including all three flavours, was as predicted by the standard solar model. This was achieved by SNO [51] by measuring neutral current reactions which can occur for all three flavours since they do not have the energy constraint of charged currents, namely the mass of the appropriate produced charged lepton. It's heavy water (D_2O) target made it sensitive to three neutrino reactions, including neutral current reactions:

- Elastic scattering on electrons: $\nu_{e,\mu,\tau} + e^- \rightarrow \nu_{e,\mu,\tau} + e^-$
- Charged current absorption of neutrinos by deuterons: $\nu_e + d \rightarrow e^- + p + p$
- Neutral current disintegration of the deuteron with a threshold of 2.2 MeV: $\nu_{e,\mu,\tau} + d \rightarrow \nu_{e,\mu,\tau} + n + p$. This reaction could not be observed in light water because of the binding energy of oxygen being larger than the solar neutrino energies.The neutron was detected by absorption on deuterium or on ^{35}Cl in added $MgCl_2$. At a later stage of the experiment an array of ^3He filled proportional tubes was added providing a highly efficient neutron detection through the reaction $^3He + n \rightarrow p + {}^3H + 764\,keV$.

Because of the overall neutron production of only a few tens per day, care had to be given to select radiopure materials. The detector is located in a mine at a depth of 6000 m.w.e, thus reducing the cosmic ray background to 70/day. It consists of an acrylic sphere containing 1000 tons of heavy water viewed by 9438 photomultipliers. It is immersed in a structure containing light water for shielding and support. The proportional counters were placed in the heavy water in a lattice with 1 m spacing. The counters were 5.08 cm in diameter and filled with 85% ^3He and 15% CF_4 at a pressure of 2.5 atm. Electrons were detected by the Cerenkov light they emitted. These included those produced in the primary interaction as well as those produced through Compton scattering on electrons of photons emitted through neutron absorption.

Cerenkov detectors are also the technique of choice for cosmological neutrinos. The scarcity of these very high energy neutrinos requires the use of large naturally occurring target and detection media such as a lake [53] or sea water [54–57] or Antarctic ice [58, 59] which can be instrumented with photomultipliers at the scale of 1 km^3. The photomultipliers are connected into vertical strings and lowered in the water or, in the case of ice, into holes melted using hot water. The strings have to be located at great depths to shield the detector from downgoing cosmic muons. This necessitates the inclusion of the photomultipliers in pressure vessels. They must also be in regions of high light transmission in order to maximize the spacing of photomultipliers and reduce the cost. The most advanced of these detectors is ICECUBE [59] in the Antarctic. It consists of 86 strings positioned in a 125 m hexagonal grid at a depth between 1450 and 2450 m. Each string includes 60 digital optical modules (DOM). Each DOM is a 35 cm pressure vessel containing a 25 cm diameter pmt, a wave form digitizer, a fast ADC and electronics self-triggering at the level of 1/4 of a photoelectron. Digital information is sent to the surface. It is complemented by a 1 km^2 surface array consisting of 160 ice-filled tanks. The average absorption and scattering lengths of the ice at the detector depth are 110 m

and 20 m respectively at 400 nm. It's energy threshold is 100 GeV. A subarray, DeepCore, consisting of 8 strings closely spaced at 40–70 m and with a DOM separation of 7 m instead of 17 m allows the observation of neutrinos with energies as low as 10 GeV. ICECUBE can search for point sources with an angular resolution of 1.5°, based on the signal arrival time at the photomultipliers, which is also used to reject downgoing cosmic ray muons. ICECUBE made the first observation of cosmological neutrinos between 20 and 2000 TeV, at energies high enough that they could not be attributed to atmospheric neutrinos. Several extensions of ICECUBE are being considered. ICECUBE-Gen2 [60] consists of an additional 120 strings to augment the coverage by about an order of magnitude, coupled with new more directional sensitive detectors as well as smaller ones to reduce the hole diameter and hence the fuel cost. Another is PINGU [61], a proposal to study neutrino oscillations parameters using a sample of about 60,000 atmospheric neutrinos with a threshold energy of a few GeV obtained by instrumenting a 6 Mton clear ice volume at the bottom of ICECUBE with 26 closely spaced strings each carrying 192 optical modules.

In addition to optical detection of the Cerenkov light, Antarctic ice has also been used to detect the coherent radio signals emitted by the cascade resulting when a neutrino interacts in a dielectric medium, the ice. This kind of radiation was predicted by Askaryan [62] and is caused by propagating showers acquiring a negative charge excess through Compton scattering and the annihilation of positrons in the dielectric. When this excess moves at a velocity greater than the velocity of light in the medium, Cerenkov radiation will be emitted and will be coherent for wavelengths longer than the transverse dimension of the shower, corresponding to $\sim$1 GHz. The electric field strength will be proportional to the shower energy. The radio attenuation length has been measured to be about 1600 m at 300 MHz, making the ice suitable for widely spaced detectors. This technique has been applied using either detectors observing the ice from balloons and satellites or with detectors placed right on the ground. The first technique allows the observation of large volumes of ice but will have higher detection thresholds. The second, due to the proximity of the detectors to the ice will have lower detection thresholds but will be limited to smaller detection volumes. The ANITA [63], Antarctic Impulsive Transient Antenna experiment, is a good example of the first technique. It used a ballooon flying under the NASA Long Duration Balloon program at an altitude of 37 km which allows the observation of the whole antarctic ice sheet (1.5×10^6 km^2). It flew 3 times for 35, 31 and 22 days respectively and used horizontal and vertical polarization antennas with a band width of 200–1200 MHz. The data was read with 2 GSamples/s digitization resulting in a 100 ns waveform per channel and per trigger. ANITA has been able to set the best limit on neutrinos for energies greater than $10^{19.5}$ eV as well as finding no neutrino coincident within 10 min of any of 12 Gamma Ray Bursts (GRBs). A possible extension of this technique would be EVA [64], the Exa Volt Antenna project, which would lower the energy threshold by a factor of 10 using the inner surface of a super-pressure balloon as a toroidal reflector antenna 115 m in diameter.

ARA [65], the Askaryan Radio Array, is a ground based radio array using several stations of 16 antennas each embeded 200 m deep into the South Pole ice. Three stations are operational with two more being installed. Each station consists of four strings separated by 10 m and each consisting of a mixture of horizontal and vertical polarization antennas. The trigger requires 3 out of 16 signal to exceed a power threshold within 110 ns. So far, the deployed stations have found no neutrino candidate including a search centred on 57 GRBs. Ground arrays have also been proposed on the Ross Ice Shelf, (ARIANNA [66]) and in Greenland (GNO [67]).

This technique has also been extended to neutrino interactions in underground Rock Salt which allows for better shielding from cosmic rays and also to interactions in the loose layer of regolith sands on the moon surface. Both of these materials have attenuation lengths for radio waves of the order of 100 m. While the rock salt experiments use ground based detectors, the lunar ones, use various radio telescopes pointed at the lunar limb.The first lunar radio experiment was the Parkes Lunar Radio Cherenkov experiment [68] and was followed by LUNASKA [69], GLUE [70] which constrained the neutrino flux above 10^{21} eV and NuMooN [71] which constrained it above 10^{23} eV.

Following the good performance of ANTARES off Toulon in the Mediterannean, a northern hemisphere kilometer cube detector, KM3Net, is currently being implemented [72]. It will consist of two modules. ORCA, at the ANTARES site will consist of 115 closely packed strings in order to address neutrino oscillations and the mass hierarchy in the energy range 3–50 GeV. The site will have a diameter of 200 m and a height of 100 m. ARCA, off Capo Passero in Sicily, will consist of two blocks of 115 widely spaced strings each block having a diameter of 1 km and a height of 600 m. ARCA's physics objectives are neutrinos from extra terrestrial sources above 1 TeV and the origin of high energy cosmic rays. Each string of both modules will consist of 18 optical modules, each housing 31 7.5 cm diameter photomultipliers. These yield a photocathode area that exceeds by a factor of three that of a single 25 cm photomultiplier, provides some directional information and a good separation between one and two photoelectron signals.

Lastly a detector,GVD [73],the Gigaton Volume Detector is under construction at Lake Baikal to observe cosmological neutrinos. It will consist of eight 120 m diameter clusters of 8 strings each, each string carrying 36 optical modules housing a $10''$ Hamamatsu photomultiplier. The cluster separation is 300 m. It is planned to extend GVD to 18 clusters to make it a cubic kilometer detector.

A detector immersed in the sea in the gulf of Taranto had also been proposed [74] to study $\nu_\mu \rightarrow \nu_e$ oscillations in the then CNGS beam [75], the axis of which was at a height of 40 km above the surface, thus placing the detector in an off-axis location. In this case the array of photomultipliers consisted of a vertical plane facing the beam to observe Cerenkov light from the electrons and muons produced in charged current interactions and identify them from their light pattern.

8.3.2.1 Liquid Argon

The Liquid Argon Time Projection Chamber (TPC) is a detector technique that provides accurate imaging of interactions while providing a moderately dense medium ($1.4\,\mathrm{g}\cdot\mathrm{cm}^{-3}$) thus very suitable for neutrino physics. It consists, Fig. 8.11, of a volume of liquid argon sandwiched between a cathode and anode providing a drift field of the order of 500 V/cm. Charged particles traversing this volume ionize it and the resulting electrons drift to the anode. To improve the uniformity of the electric field, a field cage surrounds the volume between the cathode and anode and is constructed with hoop-shaped electrodes held at potentials that increase from the cathode to the anode. The anode consists of a succession of wire planes, usually two or three. The first planes encountered by the drift electrons are biased such as to prevent them from being captured. Signals in these planes are by induction only. The last plane actually collects the electrons. The wires of the different planes are at differing orientations to the vertical. Associating signals in the different wire planes according to their timing allows the reconstruction of two of the coordinates of the drift electrons and therefore of the track portion they originated from. The third coordinate, along the drift field, is obtained using the drift velocity ($\sim$1 mm/μs), and the time difference between the electron signal at the wire and the time of the neutrino interaction. The latter is in turn obtained from the timing

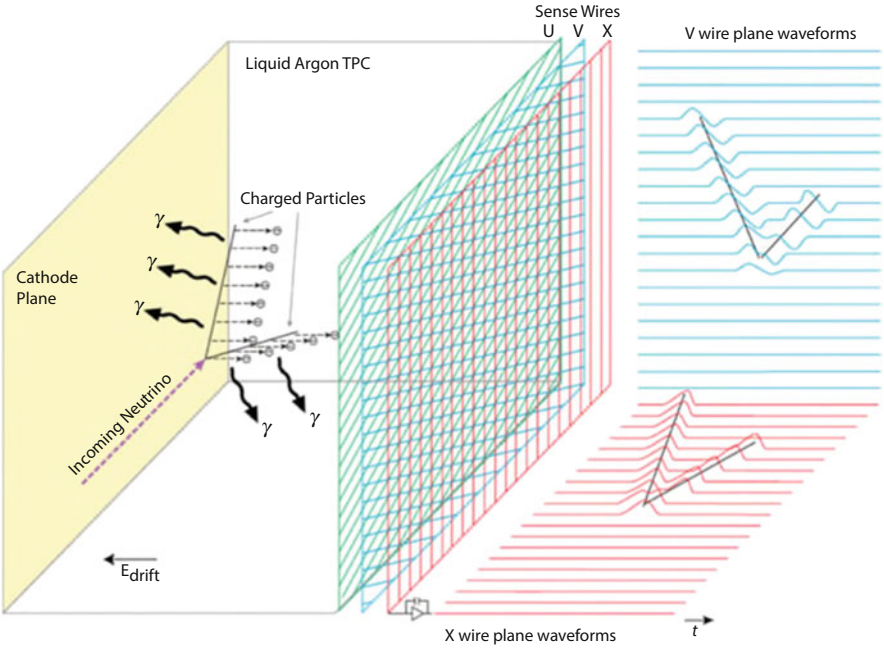


Fig. 8.11 The principle of signal recording in a Liquid Argon Time Projection Chamber as depicted in [78]

of the scintillation light emitted in the argon by the products of the interaction and recorded by photomultipliers. A track deposits energy along its trajectory and this is recorded as pulse heights in the wires. The pulse height distribution provides particle identification through the ionization pattern whereas the pulse height sum is a measure of the particle energy. The latter can also be obtained by range.

ICARUS [76], was the first to develop and use this technique. It was located at the Gran Sasso LNGS laboratory and was exposed to neutrinos produced by the CNGS beam 732 km away at CERN. It consists of two 300 ton modules each $3.6 \times 3.6 \times 19.9\,\mathrm{m}^3$. Each module includes a central high voltage plane and, along each of its long sides, three sets of detection wire planes, with orientation at 0° and ±60°. Electrons drift over a maximum distance of 1.5 m in the electric field perpendicular to the wire planes. This very complete detector relies on long drift distances and therefore requires high purity liquid argon. The purity achieved [77] during a technical run was such as to allow an electron drift lifetime of 1.8 ms equivalent to an electron mean free path of 280 cm. The electron drift velocity at 89°K increased from 0.5 mm/μs at an electric field of 0.1 kV/cm to 2 mm/μs at 1.0 kV/cm.

In the US, the first liquid argon TPC used in a physics experiment was ArgoNeuT [82], a 0.35 ton detector installed upstream of the MINOS near detector in the NUMI beam line at Fermilab. It produced significant low energy neutrino energy results as well as providing a test bed for subsequent larger liquid argon detectors.

The liquid argon technique has since been adopted for SBN [83], the Short Baseline Neutrino beam program at Fermilab, intended to investigate the possibility of additional, sterile, neutrinos. It uses the Booster Neutrino beam and consists of three liquid argon TPC detectors: SBND at 110 m from the neutrino source, MicroBooNE at 470 m and ICARUS at 600 m. The first to be installed was MicroBooNE [84], approved to observe electrons and photons and determine the origin of the low energy electromagnetic excess observed by MiniBooNE (Sect. 8.3.1). It's good spatial resolution would allow it to distinguish converting photon showers which are not associated to the primary vertex and are twice minimum ionizing at the conversion point from prompt electrons which are connected to the vertex and are singly ionizing. This should result in a good electron/photon discrimination. The TPC is inserted in a foam insulated cylindrical cryostat. It is 10.4 m long, 2.3 m high and 2.5 m wide. Electrons drift horizontally over a maximum of 2.5 m (corresponding to a maximum drift time of 2.25 ms) in a 0.273 kV/cm electric field and are recorded by two induction and one collection successive wire planes inclined respectively at ±60° and 0° to the vertical. The experiment was the first liquid argon TPC experiment to fill its cryostat without prior evacuation. It has achieved [85] an electron drift-lifetime of 18 ms corresponding to an O_2 equivalent contamination of 17 ppt and a loss of signal of 12% over the 2.5 m drift length. It also placed preamplifiers and shapers in the cold to reduce connection lengths and hence electronic noise. The amplified signals exit the cryostat and are digitized in warm ADCs before entering the DAQ electronics for Huffman compression and storage. This is done in two independent streams. The first stream records all the data occuring over

8 ms encompassing the trigger. This long recording time relative to the maximum drift time allows the complete reconstruction of cosmic rays traversing the TPC before or after the trigger but having part of their tracks reaching the wires during the event drift time. The second stream, intended for non-beam physics studies such as supernovae neutrinos, records all the data continuously but applies a zero suppression algorithm. Both the fast (6 ns) and slow (∼1μs) components of the argon ultra violet scintillation light are recorded by 35 Hamamatsu 5912-02MOD photomultipliers installed behind the anode wires and coated with Tetraphenyl Butadiene (TPB) to wave length shift the light from the ultra violet to the visible. The fast component is used to provide a trigger in time with the 1.6 μs beam spill and to tag cosmic ray tracks entering the detector during the event drift time. These are also tagged by a cosmic ray detector surrounding the cryostat and assembeled out of scintillation bars read by Kuraray WLS Y11 (200) S-type multiclad wave length shifting fibers and Hamamatsu S12825-050P multi-pixel silicon photomultipliers. A UV laser is used to map the TPC electric field, especially in the regions of non-uniformity caused by space charge effects. MicroBooNE has been collecting data since 2015. It has developed algorithms to distinguish between cosmic rays entering the detector at a rate of 4 kHz (because of its surface location) and neutrino interactions. It is also in the process of developing recontruction algorithms for electromagnetic showers that, at these low energies, can include gaps due to the propagation of low energy photons. Nonetheless liquid argon provides a remarkable visualization of events as depicted in Fig. 8.12. MicroBooNE is employing a Deep Learning technique [86] called semantic segmentation for the identification of the various classes of interactions.

The second detector to be installed will be ICARUS refurbished under the WA104/NP01 programme [87] at CERN. The following improvements were made to the detector:

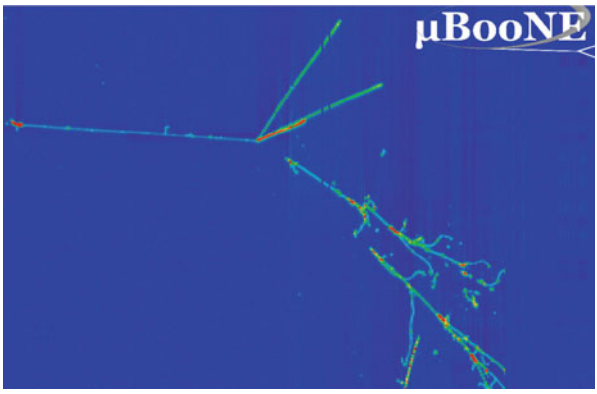


Fig. 8.12 A MicroBooNE Neutrino intearction event, showing charged particle tracks originating at the vertex and two photons, probably from a pizero, converting away from the vertex but pointing back to it

- The cathode plane was rebuilt to correct for up to 5 mm non-planarity.
- The optical system was upgraded to 360 8″ Hamamatsu 5912-mod (10 stage) cryogenic photomultipliers with TPB coating their face and read out by the CAEN V1730B 500 MHz 14 bit ADC system. The speed of this readout should allow the correlation of beam events with the Booster RF substructure, namely 1.15 ns pulses separated by 19 ns. If achieved, this correlation will reduce further the contamination of cosmic rays.
- The TPC electronics was modified as follows. The analogue signal shaping time was reduced to 1.5μs to match the electron transit time between wire planes and reduce undershoot in induction. Serial ADCs as well as a serial bus architecture with optical links were adopted. The feedthrough flange was used as the electronics backplane.

The third detector, SBND, described and referred to as LAr1-ND in [83], is a detector intended to measure the intrinsic beam composition, in particular of ν_e, before oscillations can occur. However its closeness to the neutrino target ensures a large number of neutrino interactions and hence a rich cross-section measurement programme. Its dimensions are 5 m along the beam, 4 m in height and 4 m laterally. The electrons drift along this latter dimension which consists of two 2 m drift spaces placed side by side. The Cathode Plane Assembly, CPA, is located in the middle and one Anode Plane Assembly, APA, is placed on either side and consists of the same number of wire planes and orientation as MicroBooNE. Each APA is made up of two 2.5 m wire frames along the beam but the U and V wires are connected to ensure continuous coverage. Unlike MicroBooNE, the ADCs will be in the cold together with the front end pre-amplifiers and shapers. The digitized signals will be multiplexed out of the cryostat via an FPGA. This will reduce the electronic noise and reduce the size of feed throughs. Upon exitting from the cryostat the signals will be converted to optical signals and sent, over optical fibres, to the warm DAQ electronics which will be identical to the one used by MicroBooNE. A 100 kV high voltage will provide a 500 V/cm drift field, the uniformity of which will be ensured by a field cage constructed with roll-formed tubes. A cosmic ray tagger of similar construction to the MicroBooNE one and a membrane cryostat will encase the detector. The light detection sytem will use the same pmt type and readout system as ICARUS. SBND will pioneer several detector concepts such as APAs and CPAs intended to be applied to the DUNE detector.

The liquid argon technique has been chosen for DUNE [88], the Deep Underground Neutrino Experiment $\nu_\mu \rightarrow \nu_e$ oscillation search intended to determine whether CP is violated in the neutrino sector and to measure the mass hierarchy. It will also address non-neutrino beam physics such as potential supernovae, proton decay and nnbar oscillations. The liquid argon technique was chosen instead of that of water Cerenkov for its good electron/photon discrimination resulting in a higher electron efficiency and therefore the possibility to use a smaller detector to achieve the same sensitivity. DUNE will be located 1475 m underground at SURF [89], the Sanford Underground Research Facility, in Lead, South Dakota and will be observing neutrinos produced at Fermilab 1300 km away. It will consist of four

TPC modules each containing 17,000 tons (10,000 tons fiducial volume) of liquid argon. The construction of the first module will follow the APA, CPA concept being tested in SBND, the so-called single-phase (liquid) approach. Its TPC dimensions are 12 m high, 14.5 m wide and 58 m along the beam. Three rows of APAs will be interleaved with 2 rows of CPAs, all oriented parallel to the beam. The APA-CPA horizontal separation, or drift length, will be 3.6 m, necessitating a 180 KV high voltage system for a 500 V/cm drift field. Each row of APAs consists of 25 vertically stacked pairs. Each row of pairs of facing APA-CPA is surrounded by a field cage. An APA consists of 4 wire planes separated by 4.76 mm with biases of -655 V, -365 V, 0 V and $+860$ V and orientation of $0°$, $+35.7°$, $-35.7°$ and $0°$ respectively. The wire separation is 4.7 mm. The TPC data is continuously digitized at 2 MHz by cold ADCs, serialized and transferred out of the cryostat on 12,000 high speed links per 10 kton module. They are received by Reconfigurable Computing Elements (RCEs) that buffer the raw data, zero-suppress it and pass it on to the trigger. While the zero-suppressed data is kept for non-beam physics, a second pass collects the full data set in regions of interest selected by the trigger. The photon detector system consists of light guides (2.2 m long, 83 mm wide and 6 mm thick) coated with TPB. The UV scintillation light impacting on the surface is re-emitted inside the bar at 430 nm and internally reflected in the guide to reach 12 SensL Cseries 6 mm^2 SiPMs. Ten such devices are mounted on each APA.

The second module will introduce a novel concept, the dual-phase approach first studied in [79] and tested as described in [80], in which the drifting electrons exit the liquid and are amplified in gaseous argon above the liquid. The concept is illustrated in Fig. 8.13 depicting the design of the dual phase prototype, ProtoDUNE-DP [81], currently under test at CERN. The DUNE dual phase module will consist of a 12 m wide, 12 m high and 60 m long homogeneous volume TPC. The electrons drift vertically over a maximum of 12 m in the 500 V/cm field provided by a segmented cathode at the bottom, the anode readout at the top and 60 stacked horizontal rectangular field rings. The reduction of the number of drift electrons reaching the wires due to absorption over the long drift space is compensated by the amplification in the gas. A 2 kV/cm field between a grid located just below the surface of the liquid and the Large Electron Multipliers (LEMs) charge amplification devices, causes the electrons to be extracted. The LEMs consist of a 1 mm thick printed circuit board with a micro-pattern of holes through its thickness and with one electrode on the top and one on the bottom surfaces. A 3 kV potential difference between the two electrodes results in a high field in each hole and the amplification of electrons entering them by about an order of magnitude through an avalanche process. The charge is collected in a two-dimensional x, y readout plane above the LEMs.

The technology of subsequent modules will depend on the performance of the single phase and double phase prototypes currently being built and tested at the CERN neutrino platform. DUNE also plans to use a near detector located close to the Fermilab neutrino source.

The addition of a magnetic field to a liquid argon detector would greatly enhance its capabilities. This has been tested [90] with an 11 L chamber placed in a 0.55 T magnetic field and the drifting properties were found to be preserved. However

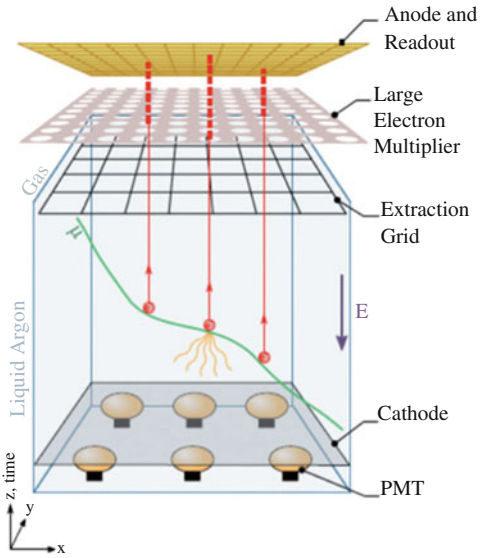


Fig. 8.13 The design of the ProtoDUNE dual phase prototype under test at CERN, showing the various components of the detector and, in particular, the amplification LEMs at the anode

measuring the charge and momentum of electrons would be challenging due to the short, 14 cm, radiation length of liquid argon resulting in only the first few cms being useful for magnetic measurements.

8.3.3 Calorimeters

These detectors are well suited to investigations requiring a measure of the total energy of an event rather than energy measurements of individual particles other than muons.

8.3.3.1 Iron-Scintillator

The CDHS detector [91] was used in the CERN SPS neutrino beam to study neutrino interactions in the energy range 30–300 GeV. It consisted of magnetized iron modules built from alternating layers of iron and scintillator and separated by drift chambers for a total of 1250 tons. Each module was constructed of circular iron plates 3.75 m in diameter and with a 30 cm central hole for the coil insertion. The coil consisted of 30 turns and was powered at 1000 A resulting in magnetic field of 1.65 T on average. It was uniform to $\pm 1.5\%$ azimuthally and dropped by about 20% with increasing radius. Two types of modules were used. Seven modules used fifteen 5 cm thick plates and twelve modules used five 15 cm thick plates. The iron plates alternated with planes of eight 45 cm wide NE110 scintillators except for

the last four modules which used a single scintillator plane for triggering. The drift chambers were 4 m wide hexagons and drifted vertically or at $\pm 60^{\circ}$ to the vertical in order to resolve ambiguities. The average efficiency was typically 99.5% and the spatial resolution 0.7 mm, which was adequate given the contribution of multiple scattering in the iron.

NuTeV/CCFR [92] used at Fermilab for a similar range of neutrino energies, differed from CDHS in that the calorimeter was separate from the magnetic spectrometer used to measure muon momenta. The 690 ton calorimeter consisted of 168 $3 \times 3 \times 5.15$ cm steel plates instrumented with Bicron 517L scintillator oil counters placed every two plates and drift chambers every four plates. This was followed by the magnetized iron toroidal spectrometer with an inner diameter of 25 cm to accomodate the four coils and an outer diameter of 350 cm. It consisted of three sections each followed by a drift chamber and two additional drift chambers downstream of the last section for improved momentum resolution. An important feature of this experiment was that a calibration beam was available in situ to determine the response [92] of the detector to electrons, muons and hadrons. The hadronic resolution was $\sigma_E/E = 0.86/\sqrt{E(GeV)} \bigoplus 0.022$ with an absolute scale uncertainty of 0.43%. The muon scale uncertainty was 0.7%, dominated by the field map determination in the iron. NuTeV performed a precise measurement of $\sin^2 \theta_W$ necessitating the measurement of both neutral and charged current events. They discriminated between the two on the basis of event length defined as the number of scintillator planes with non-zero pulse height in an event.

The 5.4 kiloton, 31 m long MINOS detector [93], similar in concept to CDHS, consists of 486 2.54 cm thick iron plates interleaved with planes of scintillator strips read by wave length shifting fibres. It was exposed to the Fermilab NuMI beam and housed in the Soudan mine 735 km away from Fermilab. This detector is studying ν_μ disappearance and therefore the shape and magnitude of the beam energy distribution must be very well understood. To minimize its dependence on Monte Carlo calculations the experiment is equipped with a near detector, located 1015 m from the target, which measures the beam spectrum and composition before any significant oscillations can occur. A transfer matrix is then used to predict the flux at Soudan. The transformation does not depend simply on the inverse of the square of the distance to the detector as the near detector, being close to the target, is not exposed to an exact point source because of the finite (725 m) length of the decay tunnel. The extruded polystyrene scintillator strips are 4.1 cm wide and 1 cm thick and are read by a 1.2 mm wave-length shifting fibre housed in a groove. The fibres are read by multi-anode photomultipliers, structured as 16 pixel in the far detector and 64 pixel in the near one. The data is multiplexed to reduce the number of readout channels. The coil provides a toroidal magnetic field in the iron allowing the measurement of the momentum and charge of secondary muons.

8.3.3.2 Fine-Grained

CHARM II [94] was a detector designed to measure $\sin^2\theta_W$ in $\nu_\mu - e$ scattering. The scattered electron is produced very forward unlike background events arising from ν-nucleon events. A cut of $E\theta^2 \leq 1$ MeV and $\theta < 10$ mrad, where E and θ are the scattered electron energy and production angle to the beam direction was used to reject this background, necessitating a very good angular resolution. Hence, glass, a low Z material to minimize multiple scattering, was selected as target material. Each of the 420 48 mm thick glass plates was followed by a plane of 352 plastic streamer tubes with a 1 cm pitch. The wires were readout in digital mode and 18 mm wide cathode strips glued to the outside of the tubes in a direction orthogonal to the wires were readout in analog mode to provide a measure of the energy and centroid of the electron showers. Consecutive modules had their strip and wire orientations rotated by 90° and consecutive modules with the same orientation had their wire spacing shifted by half the wire pitch. A scintillator plane was inserted in the detector after every 5 glass plates. The total mass of the calorimeter was 692 tons covering a volume of $3.7 \times 3.7 \times 15.4\,\mathrm{m}^3$. An electron angular resolution, σ_θ/θ, varying between 15–20 (mrad)/$\sqrt{E(GeV)}$ over the 2–24 GeV energy range of the experiment was achieved[95] as well as a vertex resolution of about 22 mm. The ability to discriminate the electrons from $\nu_\mu - e$ scattering from background is demonstrated in Fig. 8.14. A muon spectrometer consisting of six of the CDHS modules followed the glass target and provided a momentum resolution of 14% at 20 GeV/c and an angular resolution at the vertex of 18 mrad/E(GeV).

8.3.4 Emulsions

Detectors based on the photographic emulsion technique have a sub-micron spatial resolution and are therefore the detector of choice when searching for secondary vertices related to charmed particles or τ leptons. Until recently, this technique was limited because of the difficulty in scanning the emulsion. However recent developments in fast microscopes have revived it. Although ν_μ disappearance in atmospheric neutrinos was widely believed to be due to $\nu_\mu \rightarrow \nu_\tau$ interactions, it needed to be demonstrated through ν_τ appearance in a ν_μ beam. This was undertaken, using emulsions, by E531 [96] at Fermilab and by CHORUS [97] at CERN. At the neutrino energy of these experiments the τ travels only about 1–2 mm. CHORUS, the more sensitive of the two experiments used a 770 kg emulsion target built out of plates consisting of a 90 µm plastic base holding 350 µm thick emulsion layers on either side. The target was divided into four stacks each one followed by three interface emulsion sheets and a scintillating fibre tracker. Other sheets of emulsions were interleaved between the stacks and were changed several times throughout the data-taking in order to be exposed to fewer tracks and therefore ease the track reconstruction in the bulk emulsion. The fibres (more than 1 million) were read out by 58 optoelectronic readout chains each consisting of four image

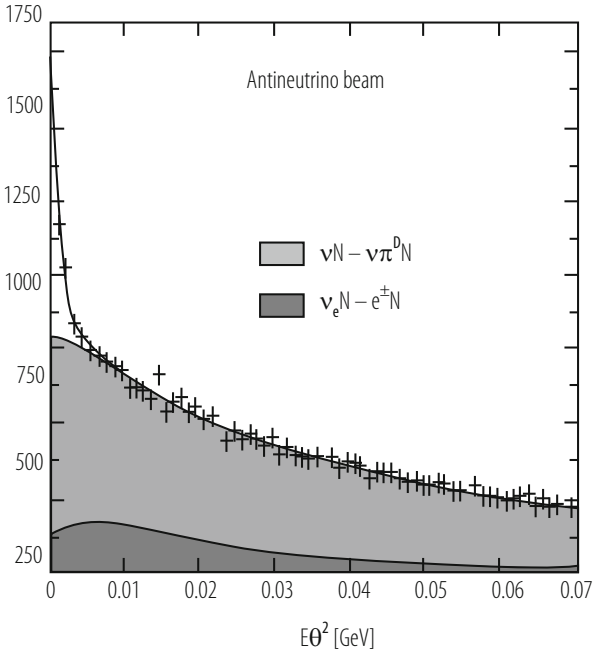


Fig. 8.14 The number of events as a function of $E\theta^2$, with E and θ respectively the scattered electron energy and direction as obtained during the antineutrino running of CHARM II, demonstrating the ability to identify electrons from $\nu_\mu - e$ scattering as evidenced by the sharp peak at small $E\theta^2$

intensifiers and a CCD camera. The target was followed by a hadron spectrometer including an air-core hexagonal magnet, an electromagnetic calorimeter and a muon spectrometer. The pulsed hexagonal magnet provided a momentum resolution varying between 20 and 50% in the momentum range 0–10 GeV/c. The scanning of the emulsions was made with fully automated Ultra Track Selector microscopes based on the track selector principle [98]. A series of tomographic images (Fig. 8.15) are taken in the emulsion at successive depths along the beam direction. Tracks then appear as aligned grains when the images are shifted according to track angle. Although they failed to find oscillations because of the kinematic region which they were sensitive to, E531 [100] and CHORUS [99] were successful in observing secondary vertices from a large sample of charm decays.

The search was then taken over by OPERA [102]. This experiment collected data at the LNGS laboratory using the CNGS beam. The long baseline of OPERA allowed the search for ν_τ appearance in the Δm^2 region then favoured by ν_μ disappearance. It used the emulsion cloud chamber technology as it is well suited to search for detached vertices or kinks over distances of the order of a mm. The 1766 ton detector was made up two supermodules. Each supermodule included 31 walls of bricks each 8.3 kg brick consisting of 57 plates of emulsions alternating

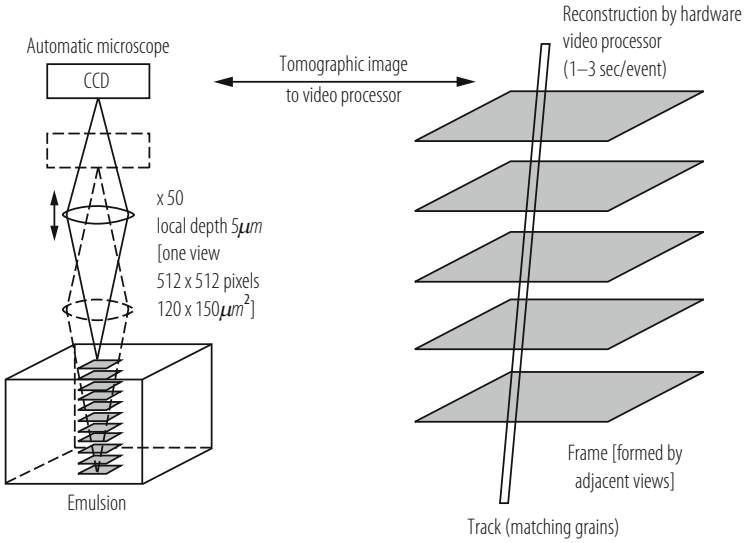


Fig. 8.15 The principle of tomographic scanning of emulsions

with 1 mm sheets of lead. An emulsion plate was made of two 44 μm emulsion sheets on either side of a 200 μm plastic layer. Each wall was followed by two planes of scintillator trackers one with vertical strips and one with horizontal strips. Each supermodule was completed with a magnetized iron muon spectrometer constructed with iron plates interleaved with resistive plate chambers and preceded and followed by 7 m long drift tubes. Bricks identified by the scintillator tracker as likely candidates for having been the site of a neutrino interaction were removed from the setup on a daily basis using a robot. They were briefly exposed to cosmic rays to provide sheet to sheet alignment and were then dismantled. The emulsions were developed and scanned. They observed [103] five ν_τ CC interactions through the detection of the resulting τ decaying to a single hadron in three events, to 3 hadrons in one event and to a muon in the fifth event. Their secondary vertex and kink detecting capability was again demonstated by their observation of neutrino generated charm events [104].

Emulsions were also used in DONUT [101], the first experiment to observe ν_τ interactions albeit from ν_τ's intrinsically present in the beam rather than from oscillations. In order to reduce the number of ν_μ and ν_e in the beam, the experiment used a neutrino beam produced by impinging the Fermilab proton beam in a 1 m long tungsten beam dump. Most pions and kaons that normally give rise to ν_μ and ν_e interacted before decaying, whereas the decay $D_s \rightarrow \tau \quad \nu_\tau$ followed by $\tau \rightarrow \nu_\tau + \ldots$ is fast enough to produce ν_τ's before the D_s interacted. This resulted in a neutrino beam with a 5% ν_τ content. Four emulsion targets were interleaved with scintillating fibre trackers. An electromagnetic calorimeter and a muon spectrometer completed the detector. Three types of emulsion targets were

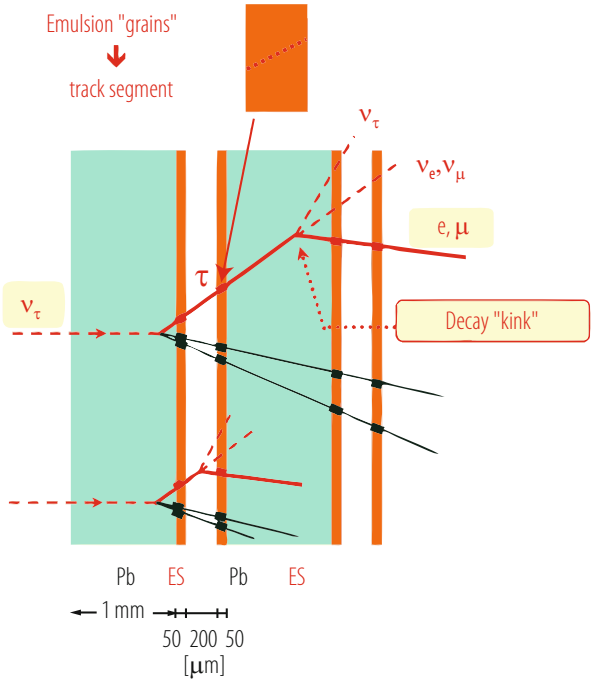


Fig. 8.16 The identification of secondary vertices using the emulsion cloud chamber technique described in the text

used. In order to increase the number of ν_τ interactions and reduce the amount of emulsion needed, in some of the targets DONUT used the emulsion cloud chamber technology, in which emulsion sheets are interleaved with lead or stainless steel plates as shown in Fig. 8.16. DONUT chose to use 1 mm thick stainless steel plates. In these targets, two types of emulsion plates were used: 100 μm emulsion sheets on either side of 200 μm or 800 μm plastic base. The remainder of the targets used bulk emulsion: 350 μm emulsion layers on either side of a 90 μm base. In the emulsion cloud chamber detectors the τ vertex is predominantly in the iron and therefore unobserved. But the precision with which the neutrino interaction vertex and the τ decay products can be reconstructed allows the identification of secondary vertices as described in Fig. 8.16. A plate to plate alignment accuracy of 0.2 μm over a $2.6 \times 2.6\,\text{mm}^2$ area was achieved by matching high momentum tracks in successive layers using position and direction information. This allowed a measure of the momentum of a particle using its multiple scattering, itself estimated using repeated changes of direction of the particle as it traverses the emulsion sheets.

This experiment also used external trackers to predict the position of interesting interactions in the emulsion. To facilitate this match each emulsion stack was followed by a changeable sheet, changed often to reduce its track density and facilitate the tracker-emulsion match. However it also used fast enough microscopes

to allow the use of stand alone techniques in which the emulsions were scanned without external information.

The latest scanning microscopes developed in Japan and Italy can measure at an average speed of 50 and 20 cm^2/h respectively. In addition to providing very precise spatial reconstruction emulsions, because of the large number of measurements along a track, can provide momentum measurements, as described above, and particle identification using ionization measurements, especially near the end point of stopping tracks. Please refer to Chap. 5 of this volume which addresses emulsion techniques in more details.

8.3.5 *Hybrid Detectors*

NOMAD [105] was a detector, Fig. 8.17, built to search for $\nu_\mu \rightarrow \nu_\tau$ oscillations by observing ν_τ interactions in the same ν_μ beam as used by CHORUS. However, unlike CHORUS, it intended to identify τ's not through the reconstruction of its separate decay vertex, but through kinematic criteria such as the missing transverse momentum arising from the unobserved neutrinos produced in τ decay. This demanded very good momentum resolution. The τ decay modes used in the analysis were $\tau^- \rightarrow \nu_\tau + hadrons$ and $\tau^- \rightarrow \nu_\tau + \bar{\nu}_e + e^-$. The latter mode was particularly useful as its main background is CC interactions of the intrinsic ν_e in the beam but this background is greatly suppressed because of the small ν_e contamination, $\sim$1%. However this mode did require very good electron identification. This latter

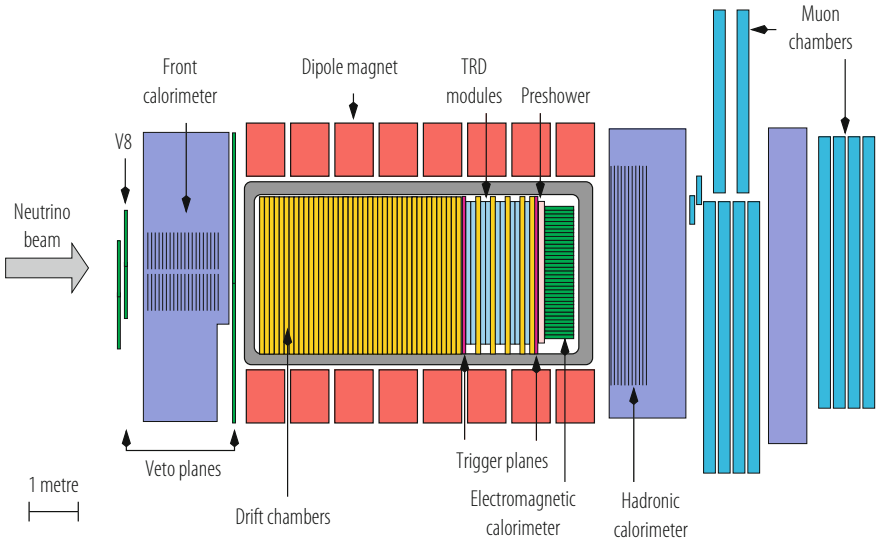


Fig. 8.17 The NOMAD hybrid detector used at CERN in the search for $\nu_\mu \rightarrow \nu_\tau$ oscillations

requirement as well as the requirement of good momentum resolution dictated the use of a light detector. The detector consisted of 49 drift chamber modules each one providing three coordinates with sense wires at 0°, +5° and −5° degrees to the vertical. The chambers were built out of honeycomb panels made of aramid fibres sandwiched between two kevlar skins. These panels provided the target material, 2.7 tons, for neutrino interactions. The average density of the target was $0.1\,\mathrm{g}\cdot\mathrm{cm}^{-3}$, close to that of a hydrogen bubble chamber and the drift chambers provided measurements every 2% of a radiation length (X_0). The spatial resolution was 150 µm providing a momentum resolution of $\sigma_p/p = \frac{0.05}{\sqrt{L}} \bigoplus \frac{0.008p}{\sqrt{L^5}}$, with the momentum p and the track length L expressed in GeV/c and meters respectively. The chambers were complemented by 9 modules of transition radiation detectors consisting of polypropylene foils and straw tubes containing an 80% xenon–20% methane gas mixture. These modules together with a 1.6 X_0 lead and proportional tube preshower and a 19 X_0 lead glass array provided the necessary e/π separation. These detectors were housed in a $7.5 \times 3.5 \times 3.5\,\mathrm{m}^3$ dipole magnet providing a 0.4 T horizontal magnetic field. The lead glass array, being inside the magnetic field, was read by tetrodes with a gain of 40 and had an energy resolution $\Delta E/E = (3.02/\sqrt{E(GeV)} + 1.04)\%$. An iron-scintillator hadron calorimeter was located just outside the magnet coil and was followed by two muon detection stations consisting of large area drift chambers located after 8 and 13 interaction lengths.

Silicon is another technique that provides very precise track localization and hence, secondary vertex identification as has been proved repeatedly in hadronic interactions. NOMAD-STAR [106] was a 45 kg prototype of a possible application of this technology to neutrino interactions. It consisted of 4 plates of boron carbide providing the interaction mass interleaved with five planes of silicon detectors. Each plane consisted of ten 72 cm long ladders of 12 silicon-strip detectors with a pitch of 50 µm. It was exposed to a neutrino beam within the NOMAD detector and, in conjunction with the rest of the detector, it was able to reconstruct 45 charm decays. The hit to noise ratio was 17:1 and the hit finding efficiency 98%. The impact parameter resolution of the μ^- produced in a ν_μ CC interaction relative to a hadronic jet consisting of at least three charged particles was 33 µm.

The magnet used by NOMAD is now being used in the T2K experiment as part of the hybrid near detector [107] located 280 m from the target. The magnet houses:

- Scintillator planes interleaved with lead or brass optimized for photon detection and π^0 reconstruction.
- Three time projection chambers (TPC) using Micromegas modules for the drift electrons amplification and readout.
- These TPC's are interleaved with fine-grained detectors consisting of strips of scintillator providing target mass.
- A scintillator and radiator electromagnetic calorimeter.
- Scintillator planes housed in the slots located in the return yoke providing a muon range detector

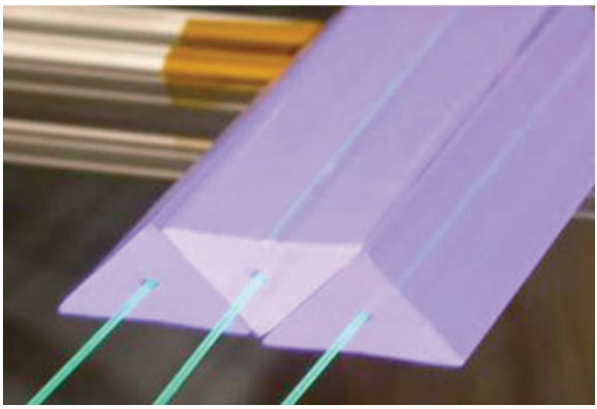


Fig. 8.18 The triangular scintillator strips and wave length shifting fibres of MINERvA

All scintillators in this near detector use Multi-Pixel Photon Counters as photosensors, a total of 50,000 channels. These are well suited as they operate in a magnetic field and provide single photon detection capability.

MINERvA, Main INjector ExpeRiment for ν-A, [108], an experiment to make precision measurements of neutrino cross sections using several nuclear targets (carbon, iron and lead) uses the NuMI beam at Fermilab and is located in front of the MINOS near detector. It consists of a fully active central detector surrounded and followed by electromagnetic and hadronic calorimeters. The central detector is built out of planes of 128 scintillator strips of triangular cross section, Fig. 8.18, and the electromagnetic and hadronic calorimeters use the lead-scintillator and steel-scintillator technology respectively. Wave-length shifting fibres are embedded in the scintillator strips and the light is channelled via clear fibres to multi-anode photomultipliers. Muons are identified and measured using the MINOS near detector. The overall cross section of the detector is hexagonal.

8.3.6 Radiochemical Detectors

Solar neutrino interactions are recorded by radiochemical experiments using the reaction: $\nu_e + (A, Z) \rightarrow e^- + (A, Z + 1)$. The atoms of $(A, Z + 1)$ produced are chemically extracted every few weeks, so this is not a real time process. They were first observed by the Homestake experiment [109] using ^{37}Cl producing ^{37}Ar. It was followed by three others, Gallex [110], Sage [111] and GNO [112], all of which used ^{71}Ga, changing to ^{71}Ge. Their characteristics are listed in Table 8.2. These experiments were housed underground to reduce cosmic ray background. In spite of the large flux of solar neutrinos on earth only a few such reactions occur

Table 8.2 Characteristics of the radiochemical solar neutrino experiments

	Homestake	Gallex	SAGE	GNO
Location	South Dakota	Gran Sasso	Baksan mine	Gran Sasso
Material	C_2Cl_4	Gallium (solution)	Gallium (metallic)	Gallium (solution)
Initial isotope	^{37}Cl	^{71}Ga	^{71}Ga	^{71}Ga
Detected isotope	^{37}Ar	^{71}Ge	^{71}Ge	^{71}Ge
Mass [tons]	615.0	30.3	57.0	30.3
Threshold	0.814 MeV	0.233 MeV	0.233 MeV	0.233 MeV
Extraction rate	3–4 months	3–4 weeks	3–4 weeks	3–4 weeks
Half-life of detection reaction	34 days	16.5 days	16.5 days	16.5 days

daily even for detectors weighing about a hundred tons due to the small neutrino interaction cross section at these low energies.

The Homestake experiment was located in the mine of the same name in South Dakota at a depth of 4200 m.w.e. The detector consisted of a cylindrical tank containing 615 tons of C_2Cl_4 and with 5% of its volume filled with helium gas at a pressure of 1.5 atmospheres. The argon produced was removed from the tank by bubbling helium through the tank and then trapping the argon in a cryogenically cooled charcoal absorber. Following several stages of purification the argon was transferred to a proportional counter after adding 7% of methane. The extraction efficiency was measured by inserting and extracting known amounts of either ^{36}Ar or ^{38}Ar. Because of the very low event rate possible radioactive contaminants in the tube material had to be minimized. The counters consisted of a highly refined iron cylindrical cathode and a 12–25 µm tungsten wire anode. The ^{37}Ar decays occur dominantly through K orbital electron capture depositing 2.82 keV of energy in the counter. This deposition is highly localized (100 µm) thus allowing it to be distinguished by pulse shape and rise time discrimination from background which is less localized.

Gallex and GNO, located in the LNGS, used 30.3 tons of Gallium containing 12 tons of ^{71}Ga in an aqueous solution acidified by the addition of HCl. This ensures that the ^{71}Ge produced is in the form of the highly volatile $GeCl_4$ in contrast with the non-volatile $GaCl_3$. The extraction procedure, as exemplified by that of GNO, is as follows [113]:

- The atoms of $GeCl_4$ (approximately 16 per 3–4 week run) are extracted into water by pumping nitrogen gas through the system.
- They are then converted into a gas, GeH_4 and mixed with Xenon.
- This mixture is introduced into proportional tubes 32 mm long and 6.4 mm in diameter made of ultrapure Suprasil quartz. The cathode consists of a single silicon crystal with impurities limited to $\leq$2ppt ^{238}U, $\leq$0.2ppt ^{232}Th and $\leq$0.1ppb ^{40}K. The anode is a 13 µm tungsten wire. The efficiency for transferring the

Germanium nuclei to the counters is measured to be 95–98% using non-radioactive Germanium carriers.
- X-rays occurring through the reactions $e^- + (A, Z+1) \rightarrow (A, Z) + \nu_e$ are detected over a period of about 6 months although the mean life of the reaction is only 16.5 days, allowing a good estimate of the background.
- The ^{71}Ge decays produce pulses of 10.4 keV or 1.1 keV for K and L captures respectively. The localized nature of this ionization allows the reduction of background using amplitude and shape analysis of the recorded pulses to a level of less than 0.1 event/day.
- The counters are calibrated 5 times during a 6-month exposure using a Gd/Ce X-ray source.

Gallex measured [114] their extraction efficiency using a 60 PBq ^{51}Cr source of 750 keV neutrinos (90%) and 430 keV (10%) neutrinos. They found a ratio of measured/expected signal of 0.93 ± 0.08. Their extraction efficiency was also confirmed [115] to be as expected to within 1% by introducing several thousand atoms of ^{71}As that decay to ^{71}Ge.

The SAGE detector was built using up to 60 tons of metallic Gallium. It was housed in the Baksan Neutrino Observatory in the Caucasus at a depth of 4700 m.w.e. While the liquid gallium was stirred at a rate of 80 rpm the Germanium was extracted from it by oxidizing it using a weakly acidic aqueous solution. The subsequent steps are similar to the procedure described above. Their extraction efficiency was also measured with a Chromium source.

The reaction threshold in chlorine only allows the observation of the beryllium and boron neutrinos whereas the threshold in gallium allows, in addition, the observation of some of the pp neutrinos.

8.3.7 Bubble Chambers

Bubble chambers were heavily used in earlier studies of neutrino interactions and were instrumental in making significant advances in the understanding of the properties of neutrinos [116]. Their filling varied from liquid hydrogen to heavy liquids, the latter used to increase the overall target mass, to contain the secondary hadrons produced and to convert photons. They were placed within a magnetic field in order to measure the momenta of the charged particles produced in the neutrino interactions. Their time resolution was poor as they were sensitive to all events occurring within a beam spill of typically millisecond duration. This could be improved by associating them to external electronic detectors.

Gargamelle was a cylindrical chamber 4.8 m long and 1.8 m in diameter. It was situated in a 2 T magnetic field produced by two coils. Neutral currents were first identified using this chamber [117] with a heavy freon (CF_3Br) filling resulting in a density of $1.5\,\mathrm{g \cdot cm^{-3}}$ and an interaction length of 58 cm. The identification of

neutral currents required the rejection of events containing muons in the final state. Muons were defined as satisfying one of the following categories:

- a particle leaving the visible volume without undergoing a nuclear scatter
- a particle which stops in the chamber and decays to an electron
- a negative particle stopping in the chamber without producing visible products (44% of negative muons are absorbed in the nucleus).

The hypothesis that the observed neutral current candidates could have been due to neutral hadrons entering the chamber was rejected as this would have resulted in a decrease of their number as a function of depth within the chamber, an effect that was not observed.

A 12-foot bubble chamber [118] (a $26\,m^3$ cylinder) was used in the neutrino beam at the Argonne Zero Gradient Synchrotron (ZGS). It was the first chamber to use a superconducting magnet.

The superconducting technology was then used in subsequent bubble chambers. BEBC (the Big European Bubble Chamber) was a $33.5\,m^3$ bubble chamber, Fig. 8.19, operating in a 3.5 T magnetic field and was exposed to the CERN neutrino beams. It was equipped with an External Muon Identifier (EMI) [119] consisting of proportional wire chambers placed behind an iron absorber and covering an area 6 m high and 25 m wide. A muon candidate track observed within the chamber was confirmed as a muon if, when its trajectory was extrapolated to the EMI, a hit was found within a distance consisting with multiple scattering. In order to reduce the miss-classification of events due to the random association of a neutral current event with a background muon in the EMI, an Internal Picket Fence [120] of proportional tubes placed between the chamber body and the magnet cryostat provided timing information for all events occurring within BEBC with a resolution of 230 ns full width. The tubes operated within the BEBC magnetic field. The chamber was used with fillings of liquid hydrogen, liquid deuterium or a neon-hydrogen mixture. This allowed the study of both νp and νn interactions. It was also used with a $3\,m^3$ Track Sensitive Target (TST) which, when filled with hydrogen within a neon-hydrogen environment provided a sample of clean interactions within the hydrogen which could be compared to interactions in the neon for which nuclear effects had to be taken into account. In addition the heavier neon allowed a more efficient detection of secondaries.

The 15-foot bubble chamber at Fermilab ran in a magnetic field of 3.0 T. When filled with a neon-hydrogen mixture [121] (61.7% atomic neon and 38.3% atomic hydrogen) it provided a 23 ton target with a density of $0.75\,g \cdot cm^{-3}$, an interaction length of 125 cm and a radiation length of 40 cm. It was also fitted with an external muon identifier.

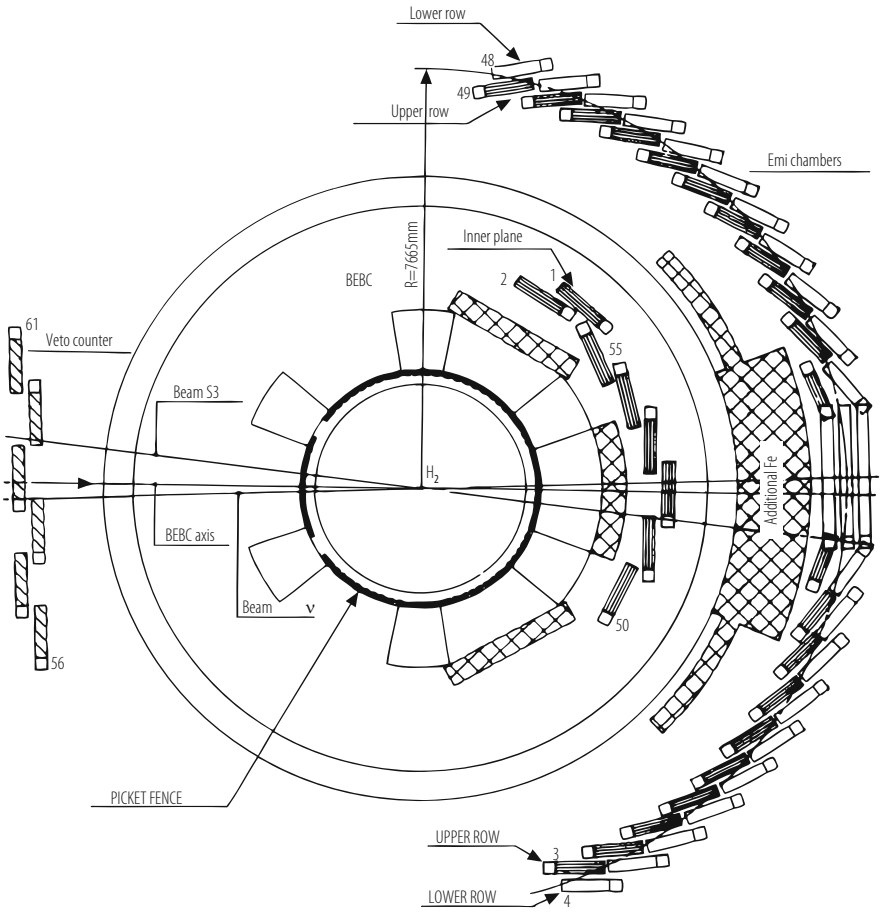


Fig. 8.19 The layout of the Big European Bubble Chamber (BEBC) including the External Muon Identifier and Internal Picket Fence

8.4 Ongoing Development Efforts on Neutrino Beams

Research based on neutrinos is currently proceeding along two paths. The first is focussed on completing our knowledge of the oscillation parameters by determining the mass hierarchy and searching for CP violation in the neutrino sector by comparing ν to $\bar{\nu}$ oscillations. The second avenue is determining whether sterile neutrinos exist. The first goal could be achieved with the presently planned detectors and beams described above. Nonetheless more accurate measurements would greatly benefit from higher intensity beams and much larger detectors. This has motivated several avenues of research pursued in the US and in Europe.

8.4.1 Beta Beams

Beta beams [122, 123] are beams of neutrinos based on the production, storage and β-decay of radioactive ions. A possible European solution was studied in the context of the Eurisol project [124]. ^{6}He ions which, decaying via β^-, produce ν_e and ^{18}Ne ions, which decaying via β^+, yield $\bar{\nu}_e$ would be stored. These ions would be accelerated to the energy required to produce decay neutrinos of the required energy and then stored in a race-track shaped storage ring. The Lorentz boost produces a well focussed forward beam which would illuminate one or more detectors in line with the straight sections of the storage ring. The search for CP violation would proceed by observing $\nu_e \rightarrow \nu_\mu$ and $\bar{\nu}_e \rightarrow \bar{\nu}_\mu$ oscillations. These oscillations would result in the observation of muons produced via the charged current interaction of ν_μ's and $\bar{\nu}_\mu$'s. These beams are particularly advantageous for the study of these oscillations as they do not, unlike present accelerator neutrino beams, have an intrinsic oscillated flavour component. Thus an important background is eliminated. The detector envisaged for this project was MEMPHYS, a water Cerenkov counter described in Sect. 8.3.2, located 130 km away from a potential beta beam source at CERN. This distance would require neutrino energies of a few hundred MeV to be at oscillation maximum. These low energies and distances would preclude any resolution of the mass hierarchy. As usual, an additional detector near the storage ring would be needed to study the beam before oscillations can occur. It is estimated that 2.9×10^{18} ^{6}He ions and 1.2×10^{18} ^{18}Ne ions decaying per year in the straight sections would be needed to meet the physics requirements. Whereas this seems achievable for ^{6}He, new production methods [125] would be needed for ^{18}Ne.

8.4.2 Neutrino Factory

A neutrino factory [126, 127] uses the decay of muons to produce neutrinos. The first step is to produce pions using a very high intensity proton beam impinging on a target. The decay of these pions then produce muons. Before they are injected into a storage ring their momentum and angular spread must be reduced to maximize their capture efficiency. This is done by phase rotation and ionization cooling. Longitudinal momentum spread would be reduced by phase rotation using the Neuffer scheme. This entails capturing multi bunches of muons with a very high Radio Frequency (RF) and rotating their phase with decreasing RF along the cooling channel. Angular spread (transverse momentum) would be reduced through ionization and subsequent longitudinal acceleration using RF cavities. The storage ring includes straight sections pointing to one or more detectors [129]. In the scheme studied in the context of the International Scoping Study [128] muons of both signs of about 20 GeV/c can be captured and stored simultaneously. Storage ring geometries have been identified that can deliver both neutrinos and antineutrinos to one or more detectors. In a race track geometry neutrinos and antineutrinos would

be identified in the same detector by time of arrival, itself related to the timing separation of μ^+ and μ^- bunches in the storage ring. In a triangular geometry two straight sections could point to two detectors. The physics envisaged with this project is the observation of $\nu_e \rightarrow \nu_\mu$ oscillations using the ν_e's produced in μ^+ decay. The signal is the observation of a μ^- as opposed to the copious μ^+'s produced by the interaction of the $\bar{\nu}_\mu$ also produced in μ^+ decay. The identification of the charge of these wrong sign muons necessitates the use of a magnetic detector. A 50 kton magnetized iron detector [130] coupled with scintillator or RPC's lends itself to this. Less dense detectors such as liquid argon TPC's and emulsion detectors [131] using the OPERA technology are also being considered to observe respectively electrons and τ leptons. These would allow the observation of additional oscillation channels which would be useful in removing ambiguities in the determination of oscillation parameters. With the higher energies and distances being considered the resolution of the mass hierarchy could be envisaged in addition to the search for CP violation.

The very high intensity proton beams needed to produce an adequate neutrino flux impose strong restrictions on the type of material used for the proton target. MERIT [132] is an R&D experiment at CERN intending to investigate the effectiveness of a mercury jet in a solenoidal field as a target. The constant flow of mercury would circumvent the problems related to stress and heating of a solid target. Muon cooling is being studied by MICE [133] at RAL with its strong synergy with MUCOOL [134] at Fermilab, with a setup including capture solenoids, liquid hydrogen absorbers and RF cavities. Incoming and outgoing spectrometers measure the effectiveness of the cooling.

8.4.3 High Current Cyclotrons

The present accelerator based long baseline experiments intend to compare oscillations of ν_μ to oscillations of $\bar{\nu}_\mu$. However the $\bar{\nu}_\mu$ beam has much more ν_μ background than the ν_μ beam has $\bar{\nu}_\mu$ background. This is because of the π^+ to π^- ratio at the proton target being larger than unity and because of the ν interaction cross section being larger than the $\bar{\nu}$ cross section. DAEδALUS [135] is an experiment aiming to remedy this situation by using the decay at rest of pions to produce a very pure source of $\bar{\nu}$ that would illuminate a detector also exposed to a long base line ν_μ beam. A beam of 800 MeV protons produced by a high current cyclotron impinges on a thick target producing pions which stop in the target, with the π^- being captured before decaying resulting in a beam dominated by the decay of π^+. As a consequence there will be essentially no $\pi^- \rightarrow \mu^- \rightarrow \mathrm{e}^- \nu_\mu \bar{\nu}_e$ and hence any $\bar{\nu}_e$ interaction observed must be from a $\bar{\nu}_\mu$ to $\bar{\nu}_e$ oscillation. The DAEδALUS project proposes installing three sources of pions at rest: one at 20 km from the detector which, for an average neutrino energy of 45 MeV, would be at the maximum oscillation probability and at the same L/E as the long baseline beam, one at 8 km to observe the rise in $\bar{\nu}_e$ appearance and one at 1.5 km for flux normalization.

The $\bar{\nu}_e$ would be observed through IBD for which a liquid argon detector as planned for DUNE would not be suitable. However, a water Cerenkov detector such as T2K or Hyper-K (especially if containing gadolinium to enhance the neutron capture rate as described in Sect. 8.3.2) or a large liquid scintillator detector would be ideal. As the flux out of the cyclotron would be continuous unlike the flux from the long baseline accelerator, the two sources of events would be distinguishable through absolute timing. The DAEδALUS collaboration is currently involved in increasing the current capability of cyclotrons to reach the 10 mA of protons necessary. The 800 MeV cyclotron would be superconducting, accelerate H_2^+ ions and would use as an injector the 60 MeV cyclotron described earlier in the context of IsoDAR. The H_2^+ ions would be stripped at extraction.

8.5 Conclusions

Neutrino detectors use the whole range of detector technologies available to high energy physicists. The smallness of neutrino cross sections necessitates the use of very large detectors that have ranged up to 50 kilotons when man-made and even 1000 megatons when using sea water or antarctic ice. The exception is the recent observation of coherent neutrino-nucleus scattering, a much larger cross section process, which allows the detection of neutrinos with smaller, albeit complex, detectors. Future generations of neutrino detectors to be used in conjunction with Very Long Base Line beams will address the oustanding questions in neutrino oscillation physics, namely the determination of the mass hierarchy and of CP violation in the neutrino sector as well as the determination of the possible existence of sterile neutrinos. In addition neutrinos are being used as probes. Ultra high energy (PeV) neutrinos originating in regions of space undergoing very violent processes are now beginning to be detected thus providing a new tool to study these processes. At the other end of the scale, neutrinos of a few MeV allow us to study the Earth and monitor reactors. These issues will require a whole range of detector sizes, up to the megatons, while at the same time requiring the precise measurements of the energies of electrons and photons and the identification of the secondary vertices of charmed particles and τ leptons. These detailed studies dictate the use of varied and complex detectors, thus ensuring that neutrino experiments will continue to use the very latest developments in detector technology.

References

1. http://www.ethbib.ethz.ch/exhibit/pauli/neutrino_e.html
2. Reines F *et al* 1953 Phys. Rev. **92** 830
3. Danby G *et al* 1962 Phys. Rev. Lett. **9** 36
4. Kodama K *et al* 2001 Phys. Lett. **B504** 218

5. Koshiba M 2008 Experimental results on neutrino masses and mixings, in Handbook of Particle Physics, Editor H. Schopper Landolt-Bornstein, Volume 1/12A
6. Camilleri L, Lisi E and Wilkerson J 2008 Neutrino Masses and Mixings: Status and Prospects Annu. Rev. Nucl. Part. Sci. **58** 343
7. Bahcall JN, Pinsonneault 2004 Phy. Rev. Lett. **92** 121301 and references therein.
8. Gaisser TK, Honda M 2002 Annu. Rev. Nucl. Part. Sci. **52** 153
9. Learned JG, Mannheim K 2000 Annu. Rev. Nucl. Part. Sci. **50** 679
10. Bemporad C *et al* 2002 Rev. Mod. Phys. **74** 297
11. Astier P *et al* 2003 Nuc. Instr. and Meth. **A515** 800
12. van der Meer S 1961 CERN Yellow Report **61-07**
13. Bernstein R *et al* 1994 Sign Selected Quadrupole Train FERMILAB-TM-1884
14. Yu J *et al* 1998 NuTeV SSQT Performance FERMILAB-TM-2040
15. Beavis D *et al* Long Baseline Neutrino Oscillation Experiment at the AGS (Proposal E889), Physics Design Report BNL 52459 (1995)
16. Dydak F 1980 Beam-Dump Experiments CERN-EP/80-204
17. Wachsmuth H 1979 Neutrino and Muon Fluxes in the CERN 400 GeV Proton Beam Dump Experiments CERN/EP 79-125
18. De Rujula A *et al* 1993 Nucl. Phys. **B405** 80
19. Apollonio M *et al* 2003 Eur. Phys. J. **C27** 331
20. Boehm F *et al* 2001 Phys. Rev. **D64** 112001
21. ExxonMobil https://ilrc.ucf.edu/documents/ILRC%2000000080/MSDS%2000000080.pdf
22. Ardellier E *et al* 2006 (Double Chooz Collaboration) hep-ex/0606025v4
23. Choi J H *et al* (RENO Collaboration) 2016 Phys. Rev. Lett. **116** 211801
24. An F P *et al* (Daya Bay Collaboration) 2016 Nucl. Instr. and Meth. **A811** 133
25. Choi J H 2016 Phys. Rev. Lett. **116** 211801
26. Abe S *et al* 2008 Phys. Rev. Lett. **100** 221803
27. Araki T *et al* 2005 Nature **436** 499
28. Alonso J R and Nakamura K (IsoDAR Collaboration) 2017 arXiv:1710.09325v1[physics.ins.det]
29. Joo K K (RENO and RENO50 Collaborations) 2017 J. Phys. Conf. Ser. **888** 012012
30. Ashenfelter J *et al* 2016 J. Phys.G:Nucl.Part.Phys. **43** 113001
31. Abreu Y *et al* 2018 arXiv:1802.02884v1[physics.ins.det]
32. Adam T *et al* (JUNO Collaboration) 2015 arXiv:1508.07166v2[physics.ins.det]
33. Learned J *et al* 2008 Hanohano: a deep ocean anti-neutrino detector for unique neutrino physics and geophysics studies. arXiv:0810.4975v1[hep-ex]
34. Alimonti G *et al* (Borexino Collaboration) 2009 Nucl. Instr. and Meth. **A600** 568
35. Derbin A and Muratova V *et al* (Borexino Collaboration) 2016 arXiv:1605.06795v1[hep-ex]
36. Bellini G *et al* 2013 arXiv:1304.7721v2[physics.ins-det]
37. Aguilar A A *et al* (MiniBooNE Collaboration) 2008 arXiv:0806.4201v1[hep-ex]
38. Aguilar A *et al* 2001 Phys. ReV. **D64** 112007
39. Ayres DS *et al* (NOνA Collaboration) 2005 arXiv:0503053[hep-ex]
40. Nitta K *et al* 2004 Nucl. Instr. and Meth. **A535** 147
41. Freedman D Z 1974 Phys. Rev. **D9** 1389
42. Akimov D *et al* 2017 arXiv:1708.01294v1[nucl-ex]
43. Bionta R *et al* 1983 Phys. Rev. Lett. **51** 27
44. Hirata KS *et al* 1988 Phys. Rev. **D38** 448
45. Fukuda S *et al* 2003 Nucl. Instr. and Meth. **A501** 418
46. Itow Y *et al* arXiv:0106019[hep-ex] and http://www2.phys.canterbury.ac.nz/~jaa53/presentations/Kato.pdf
47. de Bellephon A 2006 *et al* arXiv:0607026[hep-ex]
48. Gerigk F *et al Conceptual Design of the SPL II* CERN Yellow Report CERN 2006 -006
49. Abe K *et al* (Hyper-Kamiokande Working Group) 2011 arXiv:1109.3262v1[hep-ex]
50. The NSF multi-disciplinary initiative for a deep underground laboratory. http://www.lbl.gov/nsd/homestake

51. Boger J *et al* 2000 Nucl. Instr. and Meth. **A449** 172
52. Watanabe H *et al* 2008 arXiv:0811.0735v2[hep-ex]
53. Aynutdinov V *et al* 2008 Nucl. Instr. and Meth. **A588** 99
54. Aslanides E *et al* (ANTARES Collaboration) 1999 arXiv:9907432[astro-ph]
55. Carr J (ANTARES Collaboration) 2008 Nucl. Instr. and Meth. **A588** 80
56. Belias A (2007) in Proceedings of the First workshop on Exotic Physics with Neutrino Telescopes, EPNT06, page 97 arXiv:0701333[astro-ph]
57. Simeone F (On behalf of the NEMO Collaboration) 2008 Nucl. Instr. and Meth. **A588** 119
58. Ackermann M *et al* 2005 Astropart. Phys. **22** 339
59. Aartsen M G *et al* (IceCube Collaboration) 2017 JINST **12** P03012
60. Aartsen M G *et al* (IceCube-Gen2 Collaboration) 2014 arXiv:1412.5106v2[astro-ph.HE]
61. Aartsen M G *et al* (IceCube-Gen2 Collaboration) 2017 J.Phys. **G44** 054006
62. Askaryan G 1962 *Soviet Physics* JETP-USSR **14 (2)** 441
63. Gorham P W *et al* 2009 Astropart. Phys. **32** 10
64. Gorham P W *et al* 2011 Astropart. Phys. **35** 242
65. Allison P *et al* 2016 Phys. Rev. **D93** 082003
66. Barwick S W *et al* 2014 arXiv:1410.7369[astro-ph]
67. Wissel S A *et al* 2016 Published in PoS ICRC2015 1150
68. Hankins T H 1996 MNRAS **283** 1027
69. Bray J D *et al* 2015 Astropart. Phys.**65** 22
70. Gorham P W *et al* 2004 Phys. Rev. Lett **93** 041101
71. Buitink S *et al* 2010 Astron.Astrophys. **521** A47
72. Adrian-Martinez S *et al* (KM3Net Collaboration) 2016 J. Phys. **G43** 084001 and arXiv:1601.07459v2[astro-ph.IM]
73. Avrorin A D *et al* 2014 Nuc. Instrum. Meth. **A742** 82
74. Ball A E *et al* 2007 Eur. Phys. J. **C49** 1117
75. Acquistapace G *et al* 1998 CERN Yellow Report 98-02, INFN-AE-98-05
76. Amerio S *et al* 2004 Nucl. Instr. and Meth. **A527** 329
77. Amoruso S *et al* 2004 Nucl. Instr. and Meth. **A516** 68
78. Adams C *et al* 2018 arXiv:1802.08709v2[physics.ins-det]
79. Rubbia A 2004 arXiv:0402110[hep-ph] and http://neutrino.ethz.ch/GLACIER/
80. Badertscher A 2012 arXiv:1204.3530v3[physics.ins-det]
81. Manenti L (ProtoDUNE Collaboration) 2017 arXiv:1705.05669v2[physics.ins-det]
82. http://t692.fnal.gov/ ArgoNeut: Mini LAr TPC Exposure to Fermilab's NuMI Beam
83. Acciarri R *et al* 2015 arXiv:1503.01520v1[physics.ins-det]
84. A Proposal for a New Experiment Using the Booster and NuMI Beamlines: MicroBooNE 2007 Fermilab Proposal P974. and Acciarri R *et al* 2017 J.Inst **12** P02017
85. The MicroBooNE Collaboration 2017 MicroBooNE Public Note http://microboone.fnal.gov/wp-content/uploads/MICROBOONE-NOTE-1026-PUB.pdf
86. He K *et al* 2015 arXiv:1512.03385v1[cs.CV] and Ronneberger O *et al* 2015 arXiv:1505.04597v1[cs.CV]
87. Bonesini M (WA104 Collaboration) 2015 J. Phys.:Conf.Ser. **650** 012015
88. Acciarri R 2017 *et al* arXiv:1601.02984v1[physics.ins-det]
89. Heise J 2017 arXiv:1710.11584v1[physics.ins-det]
90. Badertscher A *et al* 2005 Nucl. Instr. and Meth. **A555** 294
91. Holder M *et al* 1978 Nucl. Instr. and Meth. **148** 235
92. Harris D A *et al* 2000 Nucl. Instr. and Meth. **A447** 377
93. Michael D G *et al* 2006 Phys. Rev. Lett. **97** 191801 and references therein
94. De Winter K *et al* 1989 Nucl. Instr. and Meth. **A278** 670
95. Geiregat D *et al* 1993 Nucl. Instr. and Meth. **A325** 92
96. Ushida N *et al* 1986 Phys. Rev. Lett. **57** 2897
97. Eskut E *et al* (1997) Nucl. Instr. and Meth. **A401** 7
98. Aoki S *et al* 1990 Nucl. Instr. and Meth. **B51** 466
99. Onengut G *et al* 2005 Phys. Lett. **B613** 105

100. Ushida N *et al* (E531 Collaboration) 1988 Phys. Lett. **B206** 375
101. Kodama K *et al* 2002 Nucl. Instr. and Meth. A **493** 45 and references therein
102. Guler M *et al* CERN SPSC 2000-028 SPSC/P318 LNGS P25/2000
103. Agafonova N *et al* (OPERA Collaboration)2015 Phys. Rev. Lett. **115** 121802
104. Agafonova N *et al* (OPERA Collaboration)2014 Eur. Phys. J. **C74** no.8 2986
105. Altegoer J *et al* 1998 Nucl. Instr. and Meth. **A404** 96
106. Ellis M, Soler FJP 2003 J. Phys. G: Nucl. Part. Phys. **29** 1975
107. Lindner T, 2008 Status of the T2K 280 m Near Detector. arXiv:0810.2220v1[hep-ex]
108. 2004 Proposal to Perform a High-Statistics Neutrino Scattering experiment Using a Fine-grained Detector in the NuMI Beam arXiv:0405002v1[hep-ex]
109. Cleveland B T *et al* 1998 Astrophys. J. **496** 505
110. Hampel W *et al* 1999 Phys. Lett. **B447** 127
111. Aburashitov J N *et al* 1999 Pys. Rv. **D60** 055801
112. Altmann M *et al* 2005 Phys. Lett. **B616** 174
113. Altmann M *et al* 2000 Phys. Lett. **B490** 16
114. Hampel W *et al* 1998 Phys. Lett. **B420** 114
115. Hampel W *et al* 1998 Phys. Lett. **B436** 158
116. Haidt D 1994 Nucl. Phys. (Proc. Suppl.) **B36** 387
117. Hasert FJ *et al* 1974 Nucl. Phys. **B73** 1
118. Barish SJ *et al* 1974 Phys.Rev.Lett. **33** 448
119. Brand C *et al* 1976 Nucl. Instrum.Meth. **136** 485
120. H. Foeth *et al*1987 Nucl. Instrum. Meth. **253** 245
121. Baker NJ *et al* 1989 Phys. Rev. **D40** 2753
122. Zucchelli P 2002 Phys. Lett.B **532** 166
123. Autin B *et al* 2003 J. Phys. G: Nucl. Part. Phys. **29** 1785
124. *European Isotope Separation Online* http://ganinfo.in2p3.fr/eurisol/
125. Rubbia C, Ferrar A, Kadi Y and Vlachoudis V 2006 Nucl. Instrum. Meth. **568** 475
126. Geer S 1998 Phys. Rev. D **57** 6989
127. De Rujula A, Gavela M B and Hernandez P 1999 Nucl. Phys. B **547** 21
128. Zisman M S (For the ISS Accelerator Working Group) 2008 J. Phys. Conf. Ser. **110** 112006
129. Abe T *et al* 2009 J. Inst. **4** T05001
130. Cervera-Villanueva A, MIND performance and prototyping, in the proceedings of the 9th International Workshop on Neutrino Factories, Superbeams and Betabeams-NuFact 07, editors O. Yasuda, C. Ohmori and N. Mondal American Institute of Physics, p 178.
131. Autiero D *et al* arXiv:0305185[hep-ph]
132. Bennett J *et al* CERN-INTC 2004-16
133. Kaplan D (For the MAP and MICE Collaborations) 2013 arXiv:1307.3891v1[physics.acc-ph]
134. Kochemirovskiy A *et al* 2016 in CNUM616-08-21 NuFact16 in Quy Nhon, Vietnam http://Vietnam.in2p3.fr/2016/nufact
135. Aberle C *et al* (The DAEδALUS Collaboration) 2013 physics.acc-ph arXiv:1307.2949v1 and Alonso J R 2016 arXiv:1611.03548v1[physics.acc-ph]

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 9
Nuclear Emulsions

Akitaka Ariga, Tomoko Ariga, Giovanni De Lellis, Antonio Ereditato, and Kimio Niwa

9.1 Introduction

Among all tracking devices used in particle physics, nuclear emulsion particle detectors feature the highest spatial resolution in measuring ionizing particle tracks. Emulsions have contributed to outstanding achievements and discoveries in particle physics. Although there was a period of decline of the emulsion technique, the interest in the technique has moved into the front line of physics research because of the advances in digital read-out by high-speed automated scanning and the continuous development of emulsion gel design. In particular, they are unsurpassed for the topological detection of short-lived particles and for specific applications in neutrino physics and other emerging fields. Indeed, a huge potential of emulsion detectors in applied research will be shown in this study. In this chapter, we will mainly focus on developments in experimental techniques for particle physics and briefly present a selection of the main experimental results.

A. Ariga · A. Ereditato
Laboratory for High Energy Physics - Albert Einstein Center for Fundamental Physics, University of Bern, Bern, Switzerland
e-mail: akitaka.ariga@lhep.unibe.ch; antonio.ereditato@cern.ch

T. Ariga
Faculty of Arts and Science, Kyushu University, Fukuoka, Japan
e-mail: tomoko.ariga@cern.ch

G. De Lellis (✉)
Dipartimento di Fisica, Napoli, Italy
e-mail: Giovanni.de.Lellis@cern.ch

K. Niwa
Graduate School of Science, Nagoya University, Nagoya, Aichi, Japan
e-mail: niwa@flab.phys.nagoya-u.ac.jp

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_9

A nuclear emulsion comprises a large number of small silver halide crystals, uniformly dispersed in gelatine. Each crystal has a typical diameter of 200 nm and works as an independent detection channel, which results in a very high detection channel density of O (10^{14}) channels/cm^3 in emulsion detectors. This makes emulsion detectors unique as particle detectors. The latest knowledge of the general photographic process is described in [1]. Herein, we discuss the detection principle of nuclear emulsions for ionizing particles.

The recent nuclear emulsion is made from silver bromide with a small fraction of iodide ($AgBr_{1-x}I_x$, x being the fraction of iodide, about a few mol%). The crystal structure of AgBr used for nuclear emulsions is face-centred cubic, and its shape is octahedral, as shown in Fig. 9.1. An AgBr crystal has a band gap of 2.684 eV. When a charged particle passes though the crystal, electrons in the valence band are transferred to the conduction band. Owing to shallow electron traps of 21–25 meV, the electrons diffuse inside the crystal until they are trapped in one of the sensitisation centres located at the surface of the crystal (electronic process). The sensitisation centre is artificially created via chemical sensitisation (e.g. sulphur-and-gold sensitisation), which is positively charged at the initial stage and works as an electron trap. The sensitisation centre, which traps an electron, is negatively charged; therefore, it attracts interstitial silver ions, which are ions migrating in the crystal lattice. The silver ion reacts with the trapped electron and forms a single silver atom ($Ag^+ + e^- \rightarrow Ag$, ionic process). The sensitisation centre is again positively charged, being ready to trap an electron. These electronic and ionic processes are repeated several times to form an aggregate of silver atoms, $Ag_{n-1} + e^- + Ag^+ \rightarrow Ag_n$, deepening its energy level. The energy level of an aggregate equal to or larger than Ag_4 is sufficiently deep to be "developable", and the sensitisation centre at this stage is called the "latent image centre". This signal is chemically amplified during the development procedure. The emulsion film is soaked in a developing solution, namely a reduction chemical. The above-mentioned

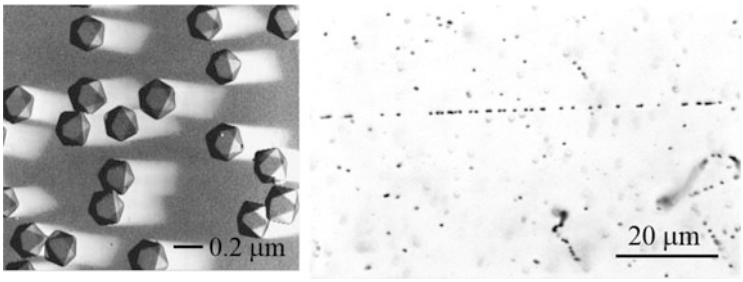


Fig. 9.1 Left: silver bromide crystals (0.2 μm linear size), as seen with an electronic microscope. Right: the track left by a minimum ionizing particle (10 GeV π^-) in nuclear emulsions; about 36 grains/100 μm are detected. Compton electrons of approximately 100 keV are also visible on right-bottom of the view

electronic and ionic processes are repeated by receiving electrons from the reducer through the latent image centre because it is a deep electron trap. This repetition lasts until all the crystals are reduced. The reaction is expressed as follows:

$$\mathrm{Red} + n\mathrm{Ag}^{+} \rightleftharpoons n\mathrm{Ag} + \mathrm{Ox} + m\mathrm{H}^{+},$$

where Red and Ox are the developing agent and the oxidized developing agent, respectively; n is the number of ions and m is the number of protons produced. Thus, a metallic silver filament remains at the position of the crystal with a latent image centre, whereas crystals without latent image centres remain unchanged. The gain of this amplification is very high, $\mathrm{O}(10^8)$. After washing out the remaining AgBr crystals via the fixing procedure, particle tracks are ready to be observed under the microscope, as shown in the right image of Fig. 9.1.

The detection efficiency of a single crystal for minimum ionizing particles (MIP) is about 0.17 [2]. The sensitivity of nuclear emulsions is translated into the number of grains per unit length. A typical emulsion has a sensitivity of 30–50 grains per 100 μm along the particle trajectory for minimum ionizing particles. Apart from the crystal size and chemical sensitisation, the sensitivity scales with the volume occupancy of the AgBr crystals with respect to the total volume of the emulsion layer, which ranges from 30 to 55%. The number of grains is proportional to the ionization power of the particle, which allows the measurement of local energy deposition (dE/dx) of each track. The random noise, so called "fog", is due to several reasons, such as thermal noise, gelatine impurity and over-sensitisation. In general, a fog density of <5 grains/10-μm-cubic is considered acceptable. In the process of producing nuclear emulsion as detectors, emulsion layers with thicknesses of 10–300 μm are formed on a glass or plastic base. To track high-energy particles (>100 MeV), a double-side coated emulsion film with a 50-μm-thick emulsion layer on either side of a 200-μm-thick plastic base is often employed. To observe both emulsion layers across the plastic base with optical microscopes, the plastic base material should not have double refraction; e.g. triacetyl cellulose and polymethylmethacrylate are appropriate.

The RMS resolution of a one-dimensional detector with a segmentation pitch of D is $D/\sqrt{12}$. Assuming that the silver halide crystal shape is approximately spherical, the resolution with a crystal diameter D is $\sqrt{\pi}D/8$ (RMS). For example, this gives 44 nm for an emulsion with 200-nm-diameter crystals. In reality, these values are slightly larger owing to the delta-ray component. A measured resolution of 50 nm (RMS) was reported for an emulsion film with a 200-nm crystal size by using high-energy particles [3], as shown in Fig. 9.2. The one-dimensional intrinsic angular resolution of a double-sided emulsion film with 200-nm-diameter crystals and a base thickness of 200 μm is therefore 0.35 mrad. Owing to the excellent position and angular resolution, one can build a vertex detector, while using a sampling calorimeter to reconstruct electromagnetic showers and also measuring the momentum of particles by the multiple Coulomb scattering, which will be discussed in Sect. 9.2. Nuclear emulsion detectors may be coupled with electronic detectors to add timing information and/or muon identification. Since emulsion

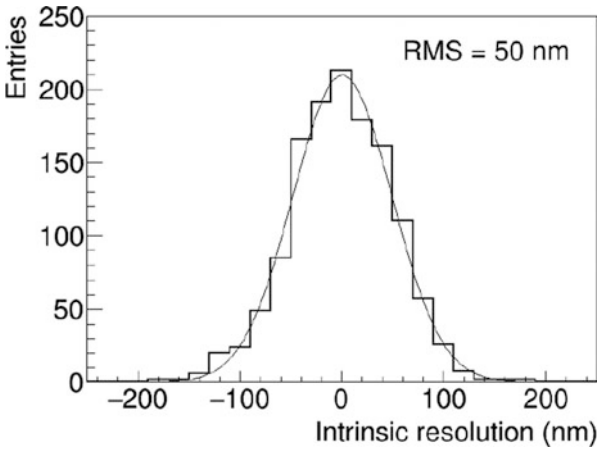


Fig. 9.2 Distribution of the distances between grains and straight-line fits to the tracks of minimum ionizing particles, showing the emulsion intrinsic spatial resolution [3]

detectors can be produced in many different sizes and shapes, there is a large variety of possibilities for a hybrid detector system, depending on physics goals, which we shall discuss in next sections. One double-sided film with a size of 10 cm $\times$ 10 cm has approximately 1-cc emulsion layer and comprises O(10^{14}) detection channels, as mentioned above. After chemical treatment, such a huge number of channels has to be read-out for physics analyses. This is the task of automated scanning microscopes, whose implementation is of fundamental importance in modern experiments making use of nuclear emulsions. This will also be discussed in the following sections.

9.2 Early Times of the Technique and the Emulsion Cloud Chamber

Thorough reviews of the basic properties and early applications of nuclear emulsions can be found in [4, 5]. The first notable examples of the use of photographic emulsion (plates) are the discovery of radioactivity by Becquerel in 1896 [6] and the measurements by Kinoshita [7], who in 1910 found records of alpha-particle radiation detected as tracks by means of optical microscopes. The emulsion technique greatly improved during the 1930s and 1940s thanks to the group of Bristol University led by Powell. He developed electron-sensitive nuclear emulsions produced by ILFORD and KODAK [8]. Powell and his group had further developed and greatly extended the seminal work of Marietta Blau. She is also known for the development of thick emulsions by a two-bath method [9].

The thickness of emulsions increased from the original 50–100 μm used in 1946 to 600–1000 μm. Even with a 500 μm thickness, a large part of the tracks of charged particles originated in the emulsion were not contained. Although an exceptional attempt to process a 2000 μm thick emulsion was reported in 1950 [10], the

difficulties of processing increased rapidly with the thickness and new difficulties appeared in the visual inspection, due to the larger scattering of light in the emulsions and the loss of optical contrast.

Plates were arranged in pairs with emulsions face to face, thus doubling the effective thickness. In 1952 a new approach was established [11–13]. Once a batch of plates was produced, the emulsions were stripped from the glass and packed together to form an almost solid sensitive mass, named stack. After exposure, the emulsions were dipped in a solution of glycerine with gelatine and then made to adhere to specially prepared glass plates. The use of a penetrating X-ray beam defined a reference frame to connect consecutive emulsion layers. With such a procedure, tracks of single particles could be quickly followed through the successive emulsions of a stack. The use of stripped emulsions became popular and allowed to make important contributions to many experiments in particle physics, as we will see in the following.

Photographic plates with 600 μm thickness were manufactured by means of newly produced emulsion gel able to record and detect the passage of ionizing particles. In parallel, dedicated microscopes were developed to observe and measure the particle tracks. With these emulsion detectors exposed to cosmic-rays Powell solved in 1947 the mystery of the Yukawa meson, by detecting the pion through its decay into a muon [14–16]. A picture of this decay as seen in nuclear emulsions is shown in Fig. 9.3. Powell was awarded the Nobel Prize for physics in 1950 for his discovery made possible by using nuclear emulsions. In the presentation of the Nobel Committee, the simplicity of the apparatus used to make such a discovery was underlined.

Few years later, in 1955, exotic hyper-nuclei were also identified by nuclear emulsions [17]. In a large balloon experiment in 1960, a 70 l emulsion chamber called "Bloury Stack" was exposed to high-altitude cosmic-rays to study their nature and the features of the induced high-energy interaction phenomena [18]. However, a crucial limitation of the technique (for those days) was met: due to the lack of scanning power, the experiment could not achieve the expected results.

A major breakthrough in the emulsion technique was the introduction of the so-called Emulsion Cloud Chamber (ECC) detector [19]. With the ECC a drastic

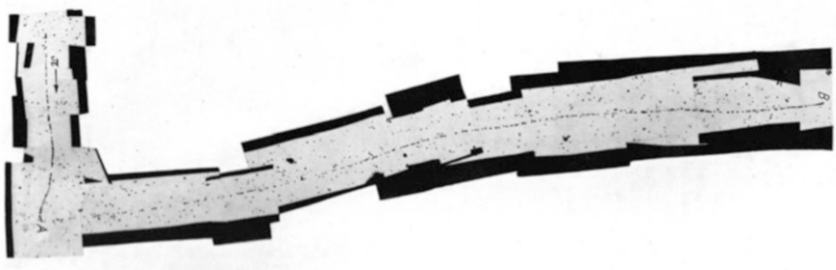


Fig. 9.3 Photomicrographs of one example of $\pi \rightarrow \mu$ decay taken from [15]

change in the detector design philosophy occurred: emulsions became a high-resolution tracking detector with three-dimensional reconstruction capabilities, rather than a visual and volume detector. This is obtained by sandwiching emulsion films or plates with passive material layers, usually made of plastic or metal plates. Today, we would call such a detector a very finely subdivided sampling-calorimeter, by means of which all charged tracks originating from the shower are reconstructed in space with high resolution. In the ECC, emulsion films are placed perpendicular to the incoming particles, so acting as a tracking detector featuring high spatial resolution (up to 1 μm).

The first design of the ECC consisted of a sandwich of brass plates and thin emulsion films. This type of detector was first developed by Kaplon and used to study heavy primaries in cosmic-ray interactions [19]. ECC detectors were applied to the study of the cosmic-ray spectrum and to very-high energy interaction processes. Nishimura, in particular, proposed the cascade shower analysis method to measure the energy of interacting γ-rays and predicted the capability of this detector to regulate the development of electron showers by selecting passive material plates on purpose [20].

Niu developed double-sided emulsion plates in which the sensitive emulsion layer is deposited on either side of a plastic substrate (see e.g. [21]). For this purpose FUJI developed a special 800 μm thick plastic base to allow gel pouring on both sides of the layer. The emulsion layers were 50 μm thick. With this new film design, two problems had to be solved: the availability of a plastic base with optical properties compatible with that of nuclear emulsions, and of a high-power objective lens with a working distance longer than 1 mm. The first problem was overcome with meta-acrylic (lucite) plates, the second was solved thanks to the efforts of Tiyoda Optical Co. The use of a plastic base between the two emulsion layers allows a precise measurement of the track angle by connecting those grains closest to the base. These points indeed are not affected by distortions. The long lever arm available with such a thick base improves the angular resolution up to 1 mrad.

The ECC opened the way to a series of important experiments of large size, thanks to the use of the dense metal plates allowing the realization of large-mass detectors with unprecedented space resolution. For the study of high-energy cosmic-rays (10 TeV) and the determination of their power law spectrum, we mention in particular the Chacaltaya experiment [22] that allowed the study of the central core of air showers, and the relatively large-size Mt. Fuji experiment [23]. For even higher cosmic-ray energies (1000 TeV and more), the RUNJOB [24] and JACEE [25] experiments studied the spectrum of primary heavy ions.

As said above, the analysis methods of ECC events are based on the reconstruction of all tracks produced following a primary interaction, likely occurring in the dense passive material. Space angles are measured for all track segments. Shower reconstruction and identification (electromagnetic or hadronic) can be performed on the basis of the topological features of the shower. In the same way, one can also reconstruct particle decays. In addition to the topological studies, powerful kinematical analyses can be conducted with ECC detectors by exploiting Multiple

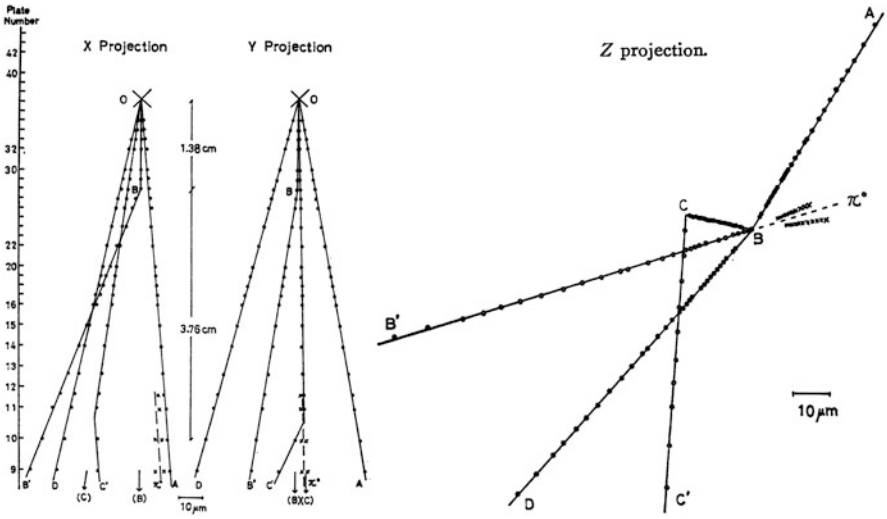


Fig. 9.4 Schematic drawing of the first evidence for the production and decay of short-lived "*X* particles" (charmed particles) in cosmic-ray interactions

Coulomb Scattering and emulsion ionization measurements, which can lead to surprisingly accurate measurements of particle momenta and particle identification.

A notable example is given by the Niu's discovery of the so-called X-particles in 1971 [21, 26]. Figure 9.4 shows a sketch of the observed topology for one event where two charged particles produced in the cosmic-ray interaction show a kink decay-topology. Today we know that this event had to be attributed to charmed meson production and decay. This happened 3 years earlier than the discovery of the J/Ψ particle by the groups of Richter [27] and Ting [28]. During those 3 years, several papers referring to that event were published by Japanese theorists [29], while there was no comparable response from the western high energy physics community. The reason for that could probably be attributed to the lack of confidence in the emulsion technique felt at that time by western scientists and also to the fact that the community carrying out cosmic-ray studies was quite apart from that employing particle accelerators, at that time not active in Japan. It is worth noting that the main distinctive features of charmed mesons were indicated already in 1974 as a result of the kinematical analysis conducted by Niu and coworkers [30], who had also re-analysed ECC data from cosmic-ray exposures carried out many years before. In addition, these authors realized that the lifetimes of charged and neutral charmed hadrons differed by a factor ranging from 2 to 3 [31].

ECC detectors allowed to design hybrid experiments combining emulsions and electronic detectors, the latter mainly used for two purposes: (1) to provide time resolution to the emulsion stack (trigger signal), (2) to pre-select the region of interest for the event occurring in the ECC, indicating the place where to start the emulsion scanning. The first hybrid experiments employed semi-automated video-

camera systems to read out the emulsion tracks and reconstruct three-dimensional vectors by measuring X, Y, θ_X, θ_Y, with Z the emulsion depth. Computers were only used to assist an operator in performing the track measurements and to provide the micro-metric movement of the microscope stage. Relative alignment was performed by fiducial X-ray marks combined with the precision measurement of the film edge positions. Typical thickness of the double-sided emulsion plate was 1 mm or larger, so allowing to follow-up tracks with a given angle $w.r.t.$ the emulsion plane by varying the focal plane of the objective lenses. The video-camera was used to grab the image from the objective lens with a rather time-consuming procedure. An operator had to manually adjust the video-image on the visually detected track, while dark spots could be automatically detected. The TV screen also allowed to run graphic tools for measuring track positions and angles.

The mechanical stability of the ECC sandwich was ensured by a vacuum packing paper known as "origami", also required to isolate the emulsion films from the external light, humidity and polluting gases. Plate-to-plate alignment was performed by X-ray lines and/or X-ray spots typically from a ^{55}Fe source. The association between the ECC and the electronic detectors was accomplished by joining particle tracks, better if of high momentum and hence less affected by Multiple Coulomb Scattering. In this respect, the idea of using interface emulsion plates in between the ECC module and the electronic detectors has proven to be very effective. These interface films were called Changeable Sheets (CS) because they were frequently replaced during the physics run in order to limit the integration of background tracks and to easily identify tracks found in the electronic detectors. This concept was first applied in the E531 experiment at Fermilab [32], as we will see in the following, and it is presently being used in several applications also for large scale ECCs.

9.3 Notable Experiments Employing Nuclear Emulsions

During the 1970s, emulsion detectors of increasing mass and complexity were developed for applications to particle physics experiments conducted at particle accelerators with experimental setups also including electronic detectors (hybrid experiments). Emulsions are often employed as active targets with high space-resolution, and the electronic detectors, namely trackers, calorimeters and spectrometers, are used to pre-select or trigger specific events in the emulsions and to complement the kinematical information of the events.

In early experiments with accelerators, nuclear emulsions were coupled to spark and bubble chambers in order to reduce the total scanning time. We recall here the observation of the decay of a charmed particle produced in a high-energy neutrino interaction in a Fermilab experiment [33]. The latter was performed in the wide-band neutrino beam produced by 400 GeV protons, by using a detector made of spark chambers placed downstream of nuclear emulsion stacks. Stacks containing altogether 16 l of ILFORD X5 emulsion made up of pellicles of 20 cm $\times$ 8 cm $\times$ 0.6 mm dimensions were placed in association with a double wide-gap spark

chamber followed by a detector of electromagnetic showers and a muon identifier. A veto counter upstream discriminated against interactions in the emulsion produced by charged particles. About 250 neutrino interactions were predicted by the spark chamber. Given its vertex position resolution, a volume of about 0.7 cm^3 was visually scanned around the prediction for about one third of the events; 16 of them were located and fully reconstructed in the emulsions and one of them was found with a topology consistent with that of charm.

A search for charmed particles in neutrino interactions was carried out at CERN in 1977 with stacks of nuclear emulsions placed in front of the entrance window of the Big European Bubble Chamber (BEBC) [34], filled with liquid hydrogen and placed in a magnetic field of 3.5 T. A veto-coincidence counter system was added in front of BEBC for this purpose. The emulsion stacks were made of 3150 pellicles of ILFORD emulsion, each 600 μm thick. The quality of the emulsion as well as the high level of muon track background precluded any systematic scanning along the track. A "surface" scan was therefore carried out for the bulk of the events with 200× and 300× objective lenses, over an emulsion volume centred on the predicted vertex position of $5 \times 31\,mm^2$ for 7 plates. A total of 206,000 BEBC pictures were analysed, leading to 935 neutrino interaction vertices inside the emulsion, 523 of which identified as charged current events. After kinematical and topological cuts, 169 charged current interactions were selected, 8 of them being identified as neutrino-induced charmed particles. The experiment reported the first direct observation of a charmed baryon decay [35] and of a neutral charmed particle [36].

The E531 experiment [32] was proposed in 1978 at Fermilab to study the properties of charmed particles and their production mechanism in neutrino interactions [37]. The neutrino beam was produced by 350 GeV protons for a first exposure (7.2×10^{18} protons on target) and by 400 GeV protons for the second one (6.8×10^{18} protons on target). The overall beam composition was 92.3% ν_μ, 7.0% $\bar{\nu}_\mu$, 0.5% ν_e and 0.2% $\bar{\nu}_e$. The active neutrino target was made of nuclear emulsions where short-lived particles were detected with micrometer accuracy. The decay products were then measured by means of an electronic spectrometer, thus making E531 the first hybrid particle physics experiment.

The emulsion target consisted of 22.6 l in the first run and of 30 l in the second one; it was made of modules composed of plates with 300 μm emulsion layers coated on both sides of 70 μm thick polystyrene foils. Downstream of the emulsion modules, two large lucite plates 800 μm thick, coated on both sides with 75 μm emulsion layers, acted as interface emulsion films, so establishing the new detector concept of the Changeable Sheets (CS). Tracks reconstructed by electronic detectors were first searched for in these interface films and then followed back in the bulk target up to the neutrino interaction vertex. The CS were replaced every 2 or 3 days of data taking in order to limit the number of accumulated background tracks that would have affected the efficiency of finding the interaction vertex in the target.

Downstream of the target, a magnet equipped with high-resolution drift chambers provided the track prediction in the CS with an accuracy of about 150 μm and 1 mrad. A time-of-flight detector made of two scintillator planes located 2.7 m

apart yielded a time resolution better than 1 ns. The setup was complemented by a lead glass array and a hadron calorimeter followed by a muon spectrometer. Three thousand eight hundred eighty-six neutrino interactions were located in the fiducial volume of the target. One hundred and twenty-two events were tagged by the presence of a secondary vertex in the target, 119 induced by neutrinos and 3 by anti-neutrinos. Events with a candidate charmed hadron in the final state were studied in detail in order to detect the presence of heavily ionizing particles (baryons) and fully reconstruct the kinematics at the decay vertex. Among those events, 57 were classified as D^0 candidates.

The analysis of the charmed hadrons is reported in [38]. Re-analyses of these results were conducted later and removed some biases present in the original studies ([39, 40]). The result on the cross-section measurements are given in [41]. In this paper, the observation of one event with the $D^0 - \bar{D}^0$ topology was reported, interpreted as associated charm production in neutral current interactions. The lifetime of charmed particles was extensively studied by E531 [42]. Limits were also set on $\nu_\mu \leftrightarrow \nu_\tau$ oscillations [43].

After the discovery of the b quark in 1977 [44], experiments with nuclear emulsions aimed at the direct observation of the production and decay of B flavored hadrons. A successful search was first performed by the WA75 experiment at CERN by using a π^- beam of 350 GeV [45]. Eight hundred and one of nuclear emulsion, in the form of double-coated plates and stripped pellicles, was exposed in 1983 and 1984. The emulsion stacks were placed both parallel and perpendicular to the beam, so exploiting the advantages of both approaches. Emulsions held perpendicular to the beam in vertical position can in fact tolerate higher track densities, while those placed parallel are more sensitive to short particle lifetimes.

The emulsion was delivered in gel form by FUJI (75 l) and ILFORD (5 l) and the pouring was done in a facility set-up at CERN [46]. Each vertical stack was made of 25 double-coated plates (330 μm thick emulsion, poured on both sides of a 70 μm thick Lexan support), $25 \times 25\,\text{cm}^2$ wide and packed in vacuum. The horizontal stacks were made of 60 stripped emulsion pellicles, 11 cm × 4 cm (4 cm along the beam) and 600 μm thick, piled-up and clamped between two rigid Perspex plates. The processing of the films was carried out in Nagoya for double-coated plates, in Rome for pellicles, and at CERN for both. After processing, each double-coated plate was cut into 64 squares, 3 × 3 cm in size, so-called mini-modules. Twenty-five squares of a module were then stuck, in sequence, on a single Lucite foil. With such a technique, the corresponding areas of consecutive emulsion plates were adjacient, thus reducing the time needed to follow a track through the stack [47]. The size of the beam was so small that it was necessary to move the target during each beam spill in order to have a uniform irradiation, thus introducing the concept of target mover. The WA75 experiment observed one event [48], schematically depicted in Fig. 9.5 as recorded in the pellicles, where both B hadrons are observed to decay into a charmed particle. The experiment also made the first observation of the purely muonic D_s decay measuring the decay constant f_{D_s} [49].

The Fermilab E653 experiment [50] was designed to measure the lifetime of B hadrons. This detector was an extension of the hybrid emulsion technique developed

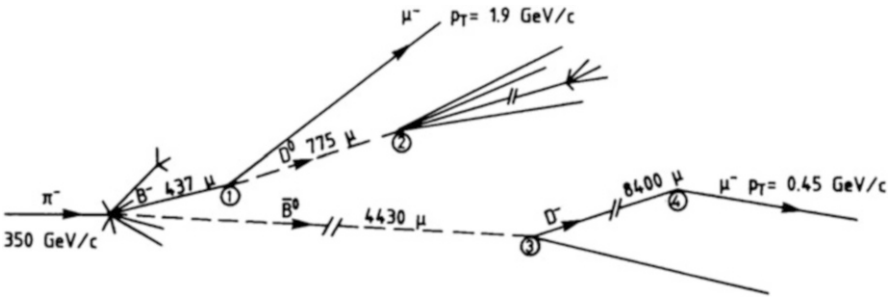


Fig. 9.5 Schematic drawing of the first hadro-produced $B\bar{B}$ pair event observed in nuclear emulsions by the WA75 experiment

for the E531 experiment and was optimized for a hadron beam. In fact, while in the E531 neutrino experiment charm was produced in one out of twenty charged current interactions, only one hadronic interaction in a thousand produces charm, and one in a million bottom. Thus, larger discrimination against non-heavy quark background was required to limit the emulsion scanning load. To achieve this, a high-resolution electronic spectrometer was placed downstream of the emulsions. Moreover, in order to cope with the large number of candidate events, the emulsion analysis required the development of computer-aided microscope techniques [51]. Events reconstructed in the spectrometer with a muon of high transverse momentum ($p_\perp > 1.5$ GeV/c) were selected for scanning in the emulsion. In the first run of 1985 a 800 GeV proton beam was used, mainly aiming at charm production. In a second run in 1987 a 600 GeV negative pion beam was exploited for the study of B mesons. Two types of target modules were employed; 55 were "vertical" and the rest "horizontal". In the first run, vertical modules were exposed to 1.5×10^5 protons/cm^2 and the horizontal ones to 0.8×10^5 protons/cm^2. The second-run exposures corresponded to 3.0×10^5 pions/cm^2 and 1.0×10^5 pions/cm^2, respectively for the two orientations.

Forty nine and fifty six target modules were exposed, respectively in the first and second run, for a total of 71 l of FUJI nuclear emulsion. Each vertical module consisted of 20 thick emulsion plates (330 μm emulsion layer on each side of a 25 cm×25 cm×70 μm polystyrene plate) and a thin film (70 μm emulsion layer on either side of a 25 cm×25 cm×500 μm lucite plate). The thin film was separated from the main block of thick plates by a 10 mm thick honeycomb, the latter combination being considered as the analysing region, while thick plates made the target region.

The emulsion modules were mounted on a target mover and displaced through the beam during the slow spill, in order to have a uniform exposure. The movement of the target was digitally controlled and the positioning encoding system granted an accuracy of 10 μm [52]. 18 silicon microstrip planes in the electronic vertex detector were located 5.7 cm downstream of the emulsion target. Secondary vertices were reconstructed by the silicon planes with typical resolutions of 6 μm transverse

to and 350 μm along the beam direction. The total fiducial decay region for bottom particles including emulsions and silicon planes was 12 cm long.

The emulsion analysis procedure first located the primary vertex. Six thousand five hundred forty-two events were selected within the fiducial volume of the emulsion and for all but 9 the primary vertex was found thanks to the excellent performance of the electronic detectors. The majority of the events were discarded by requiring a stringent angular agreement (2 mrad for tracks with a slope within 40 mrad) between the reconstructed spectrometer track and any track at the primary vertex. Three hundred and fifty-nine events in which the muon did not come from the primary vertex were retained for the secondary vertex search. Nine events met the selection criteria for bottom [53]. The *b* lifetime was also measured.

At the end of the 1980s, the production of charmed particles from quark-gluon plasma was expected to differ from that due to proton-nucleus interactions [54]. In particular, a large enhancement of the charmed quark pair creation was expected. From the experimental point of view, the major difficulty for charm detection in such nucleus-nucleus interactions came from the very short-path decay in a region close to the primary interaction where the particle density was extremely high. Two studies were carried out at CERN on this subject with emulsions, one within the NA34/2 emulsion-HELIOS programme [55] and the other one within the EMU09 Collaboration [56]. In NA34 [55], the production of charmed particles was detected in 200 GeV/nucleon ^{16}O-emulsion interactions and its cross-section was measured. Stacks of FUJI gel were exposed vertically to the ^{16}O beam. Each stack consisted of 8 double-coated plates with a surface of $25 \times 15\,\mathrm{cm}^2$ and a thickness of 700 μm (70 μm polystyrene base coated on both sides with a 315 μm thick emulsion layer).

In order to study charmed particle production in central interactions of 200 GeV per nucleon ^{32}S nuclei, the EMU09 Collaboration designed an emulsion-counter hybrid experiment at CERN [56]. The hybrid design was meant to reduce the background from secondary interactions in the emulsion, which would have spoiled the signal with heavier projectiles, differently from the case of ^{16}O. A thin and pure target was made of 100 μm thick Ag and Pb plates. Two emulsion plates, in the form of tapes, were placed downstream of the target and used as a tracking device, able to detect short-path decay vertices and producing very little secondary activity. The emulsion tape used in this experiment was derived from an Acetate base 200 μm thick and FUJI gel poured on both sides of the base, to obtain 70 μm thick layers. The emulsion analysis speed at that time did not allow to integrate a sufficiently large statistics for such rare events.

Nuclear emulsions have also played an important role in the study of multiquark systems and the quark confinement aspects of QCD. The hybrid emulsion experiment E176 [57] was carried out at KEK by using a 1.66 GeV/c K^- beam to study double-strangeness nuclei produced via Ξ^- hyperon capture at rest. Indeed, the $K^- p \rightarrow K^+ \Xi^-$ interaction produces a Ξ^-, which at rest may be captured via the process $\Xi^- p \rightarrow \Lambda\Lambda$. In the hypothesis that the H-dibaryon (ssuudd) exist, the double Λ hypernucleus can decay by emitting a H-dibaryon, in turn decaying into $\Sigma^- p$ within less than 1 mm from the Ξ^- stopping point. Unlike old-fashioned emulsion experiments where only emulsion stacks were exposed to K^- beams [58],

the hybrid design allowed the identification of the K^+ meson and the accumulation of a large statistics. Emulsion stacks were exposed vertically, perpendicular to the beam. Emulsion plates were of two types: 550 μm thick emulsion layers on both sides of a 70 μm thick polystyrene base, and 70 μm layers on both sides of a 500 μm lucite base. Thinner films with thicker base were used to avoid the degradation of the angular resolution due to distortion effects. Three double-Λ hypernuclei candidates were observed [57, 59]. However, no conclusive answer was provided on the Λ-Λ interaction. With this aim, the E373 experiment at KEK [60, 61] searched for S = -2 nuclei in nuclear emulsion with higher statistics. The apparatus was based on an emulsion-counter hybrid method, where a laser microscope performed the three-dimension graphic processing of the emulsion images, scintillating fiber blocks detected the decay products of strange particles, and a glass capillary tracker filled with liquid scintillator provided precise predictions of the Ξ^- emission angle and position. The experiment reported the observation of double hypernuclei and the Λ-Λ interaction was finally measured [62, 63]. A follow-up experiment is planned for the new J-PARC hadron facility at Tokai, still employing the hybrid detector technique with an emulsion plate stack [64].

9.4 Nuclear Emulsion Detectors with Digital Technology

9.4.1 Automated Scanning Systems and Analysis Methods

A major breakthrough in the emulsion technique occurred in 1974 when the idea of a tomographic read out of the emulsion plates was introduced by the Nagoya group [65]. In the case of an emulsion layer about 20 times thicker than the focal plane depth, one can take multiple tomographic images by sampling the emulsion layer. Those images can then be superimposed according to a given value of the presumed track slope, looking for space coincidences of the grains. After applying a detection threshold needed to remove the accidental background, a track can be defined. A first implementation of this concept led to the development of a first generation system [47] where 16 tomographic images were superimposed and a TV tube used to grab the image. This concept was developed and successfully applied to the CS emulsion scanning of the E653 experiment at Fermilab [51].

This technique was further developed by the Nagoya group and led to the so-called Track Selector [66]. The TV video was replaced by a CCD camera, yielding to higher stability and better space resolution. An FPGA-based image processor handled the 16 tomographic images of each emulsion plate. The scanning speed was actually limited by the time required for the computer-controlled objective lenses to move to the 16 different focal positions, since for each step some time was needed to damp the stage vibrations. Another limiting factor was the size of the optics field of view. A tracking efficiency as large as 90% was reached, with the main source of noise given by short Compton electron tracks. A scanning system based on a

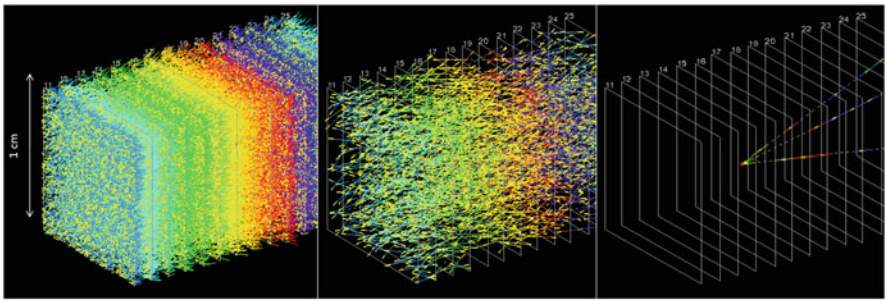


Fig. 9.6 Different steps of the emulsion data processing in the net-scan method. On the left plot all base-tracks in 15 films of the volume under study are reconstructed; they participate in the alignment process from which tracks are reconstructed, as shown in the middle plot; on the right plot passing-through tracks are discarded and the interaction vertex is reconstructed

different approach was developed in Salerno [67]. This device exploited a multi-track approach without any angular preselection.

A further important step was the establishment of fully-automatic offline analysis methods, beyond the digitization of the individual tracks around a given angle performed with the Track Selector. This progress was mainly driven by the availability of faster electronics and CCDs and more performing stage mechanics. The so-called net-scan method (described in [68]) developed in Nagoya allowed the reconstruction of tracks by associating all detected track segments regardless of their angle. Obviously, the area over which net-scan could be realistically performed depended on the available scanning speed. The latter was about 1 cm^2 per h with the UTS system [69] that exploited parallel data processing. The net-scan method allowed complete event reconstruction both at the interaction and decay vertices, precise measurements [70], search for downstream particle decays, momentum determination by Multiple Coulomb Scattering [71, 72], and electron identification by cascade shower analysis [73, 74]. Figure 9.6 shows the different steps of the emulsion data processing in the net-scan method.

9.4.2 Applications to Neutrino Experiments

In the early 1990s it became evident that the next generation of neutrino (oscillation) experiments would greatly profit from the use of the dense, high space-resolution emulsions to realize hybrid detectors well suited to the high sensitivity study of short decay topologies (charm, τ) with the possibility of a full reconstruction of the event kinematics, in turn required for background suppression. This approach was indeed motivated and justified by the advances in the emulsion technique in relation to the possibility of handling large quantities of emulsions, and also thanks to the above mentioned progress in the emulsion scanning and offline analysis, which could allow

reducing the analysis time of the emulsions by orders of magnitude as compared to the early times.

The CHORUS detector [75] is a good example of a large hybrid experimental setup combining a nuclear emulsion target with various electronic detectors. The detector was designed to search for $\nu_\mu \leftrightarrow \nu_\tau$ oscillations in the CERN WANF neutrino beam with high sensitivity. At that time, a relatively massive ν_τ was a preferred candidate to explain the Dark Matter of the Universe. Since charmed particles and the τ lepton have similar lifetimes, the detector was also well suited for the observation of the production and decay of charmed particles.

Also in CHORUS nuclear emulsions acted both as neutrino target and as a high space-resolution detector, allowing three-dimensional reconstruction of short-lived particles. The emulsion target had an unprecedented large mass of 770 kg and was segmented into four stacks, each consisting of eight modules, each in turn composed of 36 plates with a size of $36 \times 72\,\mathrm{cm}^2$. Each plate had a 90 μm plastic support coated on both sides with a 350 μm emulsion layer [76]. Each stack was followed by a set of scintillating fibre tracker planes. Three Changeable Sheets with a 90 μm emulsion layer on both sides of a 800 μm thick plastic base were used as interface between the fibre trackers and the bulk emulsion. The accuracy of the fibre tracker prediction was about 150 μm in position and 2 mrad in the track angle. The electronic detectors downstream of the emulsion target and the associated trackers included a hadron spectrometer measuring the bending of charged particles in an air-core magnet, a calorimeter where the energy and direction of showers were measured and a muon spectrometer.

CHORUS represents a milestone in the history of nuclear emulsions for the size of the target and of the CS, which implied very labor-intensive procedures for emulsion gel production, pouring on the plastic bases, and development conducted in the CERN emulsion laboratory [46], as well as for the first massive use of automated scanning microscopes running in the Japanese and European laboratories of the Collaboration [75].

The operation of the experiment consisted of several steps. It is worth noting that the large-size emulsion target was replaced only once during the entire duration of the experiment, while the CSs were periodically exchanged with new detectors, therefore integrating tracks for a relatively short period. The best time resolution was obviously provided by the electronic detectors. With the CS scanning, the association between electronic detectors and emulsions took place, and tracks with position and angle compatible with that of the electronic trackers' predictions were searched for in the interface emulsions. If found, these tracks were further extrapolated into the bulk emulsion, with a much better resolution, up to the track stopping point, with a procedure called scan-back, consisting in connecting emulsion layers progressively more upstream. After that, a "volume scan" (net-scan) around the presumed vertex was accomplished and repeated for all stopping tracks until the neutrino interaction vertex was found.

In the search for charmed particle decays, a dedicated topological selection was applied to the collected net-scan data. The analysis procedure was complemented by the visual inspection of the selected event candidates, aimed at checking both

primary and secondary vertices making used of the "stack" configuration. Decay topologies could be well separated from ordinary nuclear interactions, since the latter usually exhibit fragments from nuclear break-up or so-called "blobs" from nuclear recoil.

More than 100,000 neutrino interactions were located in CHORUS. The search for oscillations was negative and an upper limit to the oscillation probability was eventually set [77]. CHORUS reported the first observation of the associated charm production in charged current interactions [78]. This first observed event is shown schematically in Fig. 9.7. It represents the production of two charmed particles in a charged current interaction induced by a muon neutrino. Apart from six tracks of high ionization coming from the nuclear break-up and not drawn in the sketch, at the primary neutrino interaction vertex there are two charged tracks: one is the negative muon and the other one, indicated as particle 1, is a charmed hadron. The charged charmed particle shows a 417 mrad kink angle after travelling 1010 μm. The outgoing particle, indicated as particle 2, shows a flight length of 7560 μm and a reinteraction with an outgoing particle (particle 3) of high ionization. In addition to the charged charmed hadron, the decay of a neutral charmed particle is visible 340 μm downstream of the primary vertex. Two particles are generated from the neutral particle decay. The non-planarity of parent and daughter particles rules out the two-body decay and thus both the K_s^0 and the Λ hypotheses for the neutral particle. A kinematical analysis confirmed the interpretation of the event as the associated charm production induced by a muon neutrino in a charged current interaction [78]. An unprecedented statistics of about 2000 fully reconstructed neutrino-induced charmed hadron event vertices was collected. With this statistics, CHORUS measured the Λ_c and D^0 exclusive production cross-section [79] and the double-charm production cross-section in both neutral and charged current interactions [80]. The CHORUS emulsion data also provided an upper limit to the production of charmed pentaquark states [81].

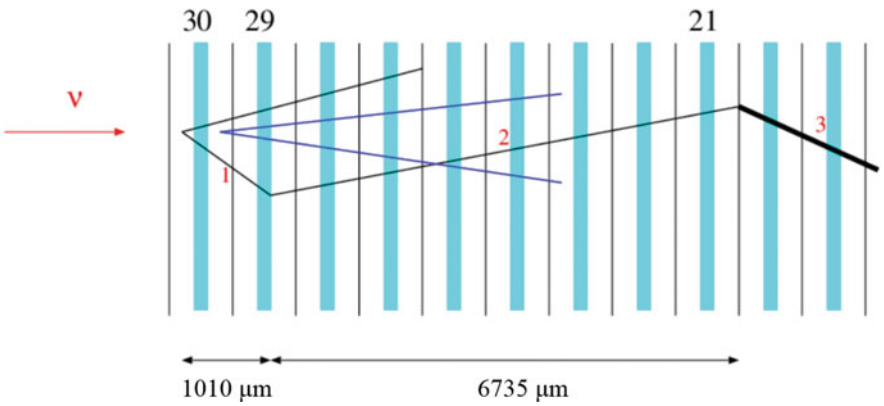


Fig. 9.7 Schematic drawing of the first neutrino-induced associated charm production event observed in the emulsions of the CHORUS experiment (see the text for explanation)

Higher sensitivity follow-ups of the CHORUS experiment were proposed, with the purpose of increasing by more than one order of magnitude its sensitivity in the measurement of the oscillations (smaller mixing angle). We mention in particular the COSMOS proposal at Fermilab [82]. The use of emulsions as large-surface trackers for the high-resolution measurement of hadron and muon momenta was proposed in [83] and then applied for the proposal of the TOSCA experiment at CERN [84]. Eventually, all those experiments were not realized mainly due to the first strong indications for $\nu_\mu \leftrightarrow \nu_\tau$ oscillations detected with atmospheric neutrinos, in disappearance mode, in a complementary region of the oscillation parameters.

The DONUT experiment at Fermilab aimed at the first direct detection of ν_τs, in this case promptly produced in a 800 GeV proton beam dump and not coming from the possible oscillation mechanism as in CHORUS. The experimental apparatus and the detection techniques used in the experiment are described in [68, 85]. The DONUT Collaboration employed an iron/emulsion ECC target able to offer a sufficiently high mass to the interaction of the neutrinos and to provide the detection of the interaction vertex, as well as a clear observation of the short track of the τ lepton (up to a few mm) produced in the ν_τ charged current interaction. The ECC was complemented by high-precision fiber trackers to drive the scan back in the emulsions.

The emulsion target eventually integrated a relatively high muon background. In a first analysis, 203 neutrino interactions were located in the ECC target, observing 4 ν_τ candidate events with an estimated background of 0.34 events [86]. This represents the first direct detection of the ν_τ. Figure 9.8 shows a display of two candidate events. In the final analysis, 9 ν_τ charged-current (CC) events were detected, with an estimated background of 1.5 events, from a total of 578 observed neutrino interactions and were used to estimate ν_τ CC cross section for the first time [87]. The main source of error in measuring the ν_τ cross section was due to the systematic uncertainties, whereas 33% of the relative uncertainty was due to the limited number of detected ν_τ events. Owing to the

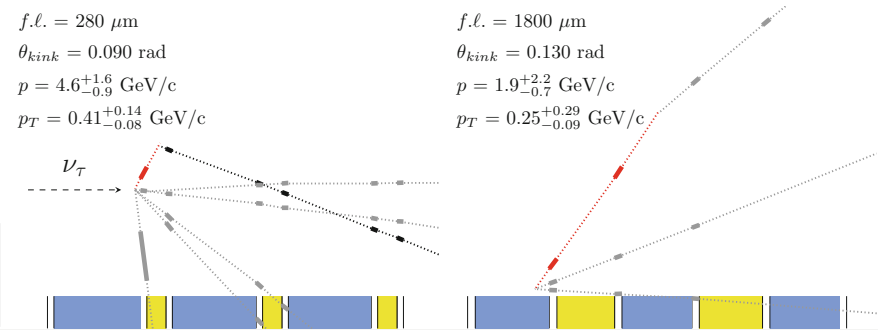


Fig. 9.8 Schematic drawing of two ν_τ induced events measured by the DONUT experiment. The kinks relative to the τ decay are visible

lack of accurate measurements of the D_s differential production cross section, DONUT expressed its ν_τ cross-section measurement as a function of the parameter n, responsible for the differential production cross section of D_s, as $\sigma_{\nu_\tau}^{\text{const}} = 2.51n^{1.52} \times 10^{-40}\,\text{cm}^2\,\text{GeV}^{-1}$. The cross section was estimated to be $\sigma_{\nu\tau}^{const} = (0.39 \pm 0.13(\text{stat.}) \pm 0.13(\text{syst.})) \times 10^{-38}\,\text{cm}^2\,\text{GeV}^{-1}$, when assuming the value of the parameter n as derived from PYTHIA 6.1 simulations.

9.5 Present Emulsion Detectors

9.5.1 Fast Scanning Systems and Large-Scale Film Production

As stated above, the advances in the scanning systems aimed at higher efficiency and speed have led in recent times to the rebirth of the emulsion detectors. A further generation of the Track Selector, called S-UTS (Super-Ultra Track Selector), was developed in Nagoya [88]. It is based on highly customized components. The main feature of this approach is the removal of the stop-and-go process of the stage in the image data taking, which is the mechanical bottleneck of traditional systems. To avoid the stop, the objective lens moves at the same constant speed of the stage while moving also along the vertical axis and grabbing images with a very fast CCD camera running at 3000 Hz. The optical system is driven by a piezoelectric device. The camera has a sensor with 512×512 pixels that imposes a smaller field of view ($\sim 120 \times 120\,\mu\text{m}^2$) to ensure a comparable position resolution (about 0.3 μm/pixel). The high-speed camera provides a data rate of 1.3 GB/s. This is handled by a front end image processor that makes the zero-suppression and the pixel packing, reducing the rate to 150–300 MB/s. A dedicated processing board performs track recognition, builds micro-tracks and stores them in a temporary device with a rate of 2–10 MB/s. A computer algorithm links the micro-tracks of different emulsion layers and writes the resulting tracks in a database that is used as input for physics analysis. The routine scanning speed is $20\,\text{cm}^2$/h/layer while one of the S-UTS systems has reached the speed of $72\,\text{cm}^2$/h/layer by using a larger field of view, without deteriorating the intrinsic micrometric accuracy of the emulsion films.

In the framework of the OPERA experiment (see next section), a joint effort of several European laboratories allowed the development of an automated scanning system (ESS) that employs commercial subsystems in a software-based framework. The ESS, derived from a system developed in Salerno [67], is extensively described elsewhere [89–91]. The microscope is a Cartesian robot, holding the emulsion film on a horizontal stage movable in $X - Y$ coordinates, with a CMOS camera mounted on the optical axis (Z), along which it can be moved to change the focal plane with a step roughly equal to the focal depth of about 3 μm. The control workstation hosts a motion control unit that directs the stage to span the area to be scanned and drives the camera along the Z axis to produce optical tomographic image sequences (with the $X - Y$ stage holding steady). Areas larger than a single field of view (∼300

$\times$ 400 μm^2) are scanned by repeating the data acquisition sequence on a grid of adjacent fields of view. The stage is moved to the desired position and the images are grabbed after it stops, with a stop-and-go algorithm. The images are grabbed by a Mpixel camera at the speed of 376 frames per second while the camera is moving in the Z direction. The whole system can work at a sustained speed of 20 cm^2/h/layer, 24 h/day, with an average data rate as large as 4 GB/day/microscope still preserving the intrinsic emulsion accuracy. A different setup of this system makes no use of immersion oil as interface between the objective lens and the film being scanned [92].

The track building method applied in both systems is schematically drawn in Fig. 9.9. The whole emulsion thickness is spanned by adjusting the focal plane of the objective lens and a sequence of 16 tomographic images is taken for each field of view at equally spaced depth levels, matching the focal depth of the objective. Emulsion images are then digitized, converted into a grey scale of 256 levels, sent to a vision processor board and analyzed to recognize sequences of aligned grains, i.e. clusters of dark pixels of given shape and size. Some of these spots are track grains; others, in fact the majority, are fog grains not associated to particle tracks. The three-dimensional structure of a track in an emulsion layer (microtrack) is reconstructed by combining clusters belonging to images at different levels and searching for geometrical alignments (Fig. 9.9a). Each microtrack pair is finally connected across the plastic base to form the so-called base track (Fig. 9.9b).

Figure 9.10 shows a S-UTS system and the scanning station in Bern employing the ESS system with dry objectives and with an automated emulsion film changer. The latter device allows fully unattended operation [93].

A second feature that significantly contributed to the rebirth of the emulsion detectors in recent times has been the realization of industrial emulsion films, optimized for micro-tracking applications. This is in particular the case of the FUJI R&D work conducted in collaboration with the Nagoya University [2] for the OPERA experiment that will be described later. Uniform automated machine

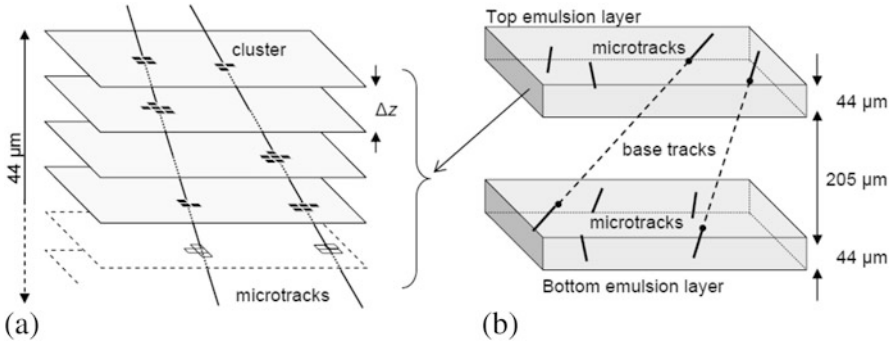


Fig. 9.9 (**a**) Microtrack reconstruction in one emulsion layer by combining clusters belonging to images at different levels; (**b**) microtrack connections across the plastic base to form base tracks

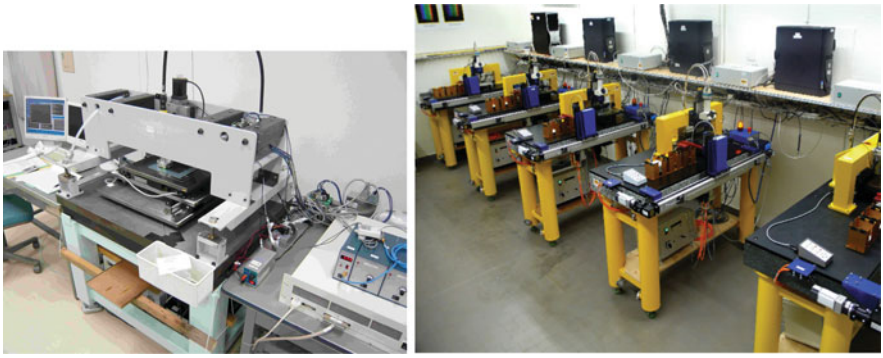


Fig. 9.10 Left: photograph of one of the Nagoya S-UTS scanning systems; right: the Bern scanning station equipped with five ESS microscopes with the associated automated film changers

coating of 44 μm emulsion layers on either side of a plastic base was achieved for the unprecedented large-scale application of the OPERA ECC modules. The quality and the uniformity of the films is remarkable.

For machine coating in an industrial plant, dilution of the gel is required in order to reduce the viscosity. This implies a reduction of the grain density. In order to recover for this degradation, improvements in the gel sensitivity were applied, such as a controlled double jet method for the production of mono-dispersion of AgBr micro crystals. The crystal size is well controlled by this method. The number of crystals along a particle trajectory is increased, while the volume occupancy of AgBr and the average diameter of the crystals is kept constant.

In order to match the experimental requirements of a relatively thick layer with the limitations coming from the industrial process, a multi-coating method was adopted by FUJI. After the first layer (20 μm thick) is coated on both sides of a rolled plastic base, a second layer is coated over the first one. A thin 2 μm gelatine spacer protects the emulsion layers. The resulting thickness is 44 μm, well sufficient for automated track recognition. A glycerine bath is used to restore the thickness of the emulsion layers to its original value, thus recovering for the shrinkage induced by the development process.

Another notable development related to the OPERA experiment has been the realization of the so-called emulsion refreshment. High temperature and high relative humidity enhance the latent image fading. This possibility is in particular useful when the exposure occurs much later than the film production and a low background is required, as in the case of OPERA. A good tuning of the fading features was achieved by introducing 5-methylbenzotriazole into the emulsion gel [2]. Absorption of this chemical by the silver specks induced by radiation lowers the oxidation reduction potential and makes the specks easy to oxidize. On the other hand, the sensitized centers (sulfur and gold) remain stable against the oxidation. Therefore, the recorded tracks are erased while the sensitivity remains sufficiently high. For example, keeping the films for 3 days at 98% relative humidity and 27 °C,

the grain density of tracks accumulated before the refreshing goes from 30 to less than 10 grains/100 μm, thus erasing about 96% of the stored tracks, including those from Compton electrons and cosmic-rays. The industrially produced films also feature a rather low track distortion induced by the development, as well as a limited level of fog density, with an initial value of 2.9 fog grains/1000 μm^3).

9.5.2 The OPERA Experiment

The OPERA experiment was designed to unambiguously prove $\nu_\mu \rightarrow \nu_\tau$ oscillations in appearance mode. Indeed, studies of atmospheric neutrinos had shown the disappearance of muon neutrinos [94], later confirmed by accelerator experiments [95] and interpreted in terms of $\nu_\mu \rightarrow \nu_\tau$ oscillations. Therefore, the appearance of tau neutrinos in a pure muon neutrino beam was the missing tile in the coherent scenario of neutrino mixing.

The conceptual design of the experiment was originally proposed in [96–98] and the detector is extensively described in [99, 100]. The distinctive feature of ν_τ charged-current interactions is the production of a short-lived τ lepton ($c\tau = 87\ \mu m$). Thus, one has to accomplish the very difficult task of detecting sub-millimeter τ decay topologies out of a huge background of ν_μ reactions in a target of more than a *kiloton*, as required to have a sufficient interaction rate. This is achieved in OPERA by employing a modern version of the ECC technology.

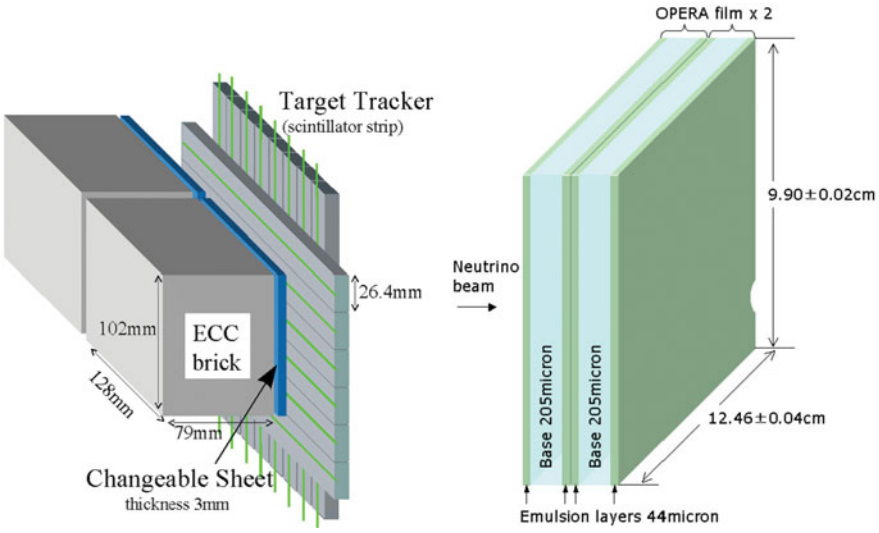


Fig. 9.11 Schematic view of the ECC unit (brick) used in the OPERA experiment. A detail of the Changeable Sheet doublet is also shown

The OPERA experiment has been running from 2008 to 2012 at the underground LNGS laboratory in Italy, 730 km away from CERN where the CNGS neutrino beam was produced. OPERA is the first very large scale emulsion experiment, profiting from all the technological advances in the emulsion technology and in the scanning systems described in the previous section. To give a figure, the ECC target is made of films with a total surface of 110,000 m^2 and 105,000 m^2 lead plates. The industrially produced, machine-coated emulsion films by FUJI provided very uniform layer thickness and the possibility of erasing unwanted background tracks by the refreshing technique. The scanning of the events was performed with about 40 fully automated microscopes, each of them faster by about two orders of magnitude than those used in the CHORUS experiment [75].

The ECC target consisted of multi-layer arrays of target walls interleaved with pairs of planes of plastic scintillator strips. A target wall (with about $10 \times 10\,m^2$ cross-section) was an assembly of horizontal trays each loaded with ECC target units called bricks. A brick consisted of 57 emulsion films interleaved with 56 lead plates, 1 mm thick, light-tight packed. Brick dimensions were $128 \times 102 \times 79\,mm^3$ for a weight of 8.3 kg (Fig. 9.11). Interface Changeable Sheets (CS) were attached to the downstream face of each brick. The choice of the CS geometry was such to assemble two adjacent emulsion films as a doublet, coupled as an independent, detachable package to the downstream face of the brick (Fig. 9.11). The use of doublets provided the cancellation of random coincidences of tracks accumulated during the storage and transportation and unerased by the refreshing procedure.

There were 150,000 bricks in total for a target mass of 1.25 kton. This represents the largest ever ECC detector assembly and posed an unprecedented challenge for the production of emulsion films and bricks, as well as for the emulsion handling, development and analysis, i.e. scanning power. Just to give some numbers, more than nine million emulsion films were produced and the corresponding 150,000 bricks were built by a fully robotised chain assembling films and lead plates in an underground dark-room at LNGS. Large infrastructures were also realized at LNGS for brick manipulation (automatic extraction from the target matrix), X-ray marking, cosmic-ray exposure and emulsion development [100].

The principle of the experiment can be summarized as follows. At the occurrence of a neutrino interaction, the resulting charged particle tracks are detected by the scintillator counter planes placed behind each brick target wall, similarly to what happens in a sampling calorimeter. The reconstruction of the "shower axis" or the identification of a penetrating track (e.g. a muon) allows identifying the brick where the neutrino likely interacted. At this point, the brick is extracted from the wall, the attached CS doublet is removed and developed, while the brick, still packed, is placed in an underground storage area waiting for the response of the CS scanning.

It is important to stress the key roles accomplished by the CS in OPERA [101]: the first step is to confirm that the ECC brick contains the neutrino interaction; the second step is to provide event-related tracks to be used for the ECC scan-back analysis. By using Compton electrons from environmental radioactivity, the systematic uncertainties in the relative alignment between the two CS doublet films are reduced, thus bringing the position accuracy to the level of 1 μm [102].

Such an accuracy allows using CS tracks made of only 3 out of the possible 4 track segments, thus increasing the track finding efficiency. Thanks to the CS, the bricks wrongly identified by the scintillator trackers are not disassembled but put back in the target with a fresh CS attached to them. This avoids useless film handling, processing and scanning of the misidentified bricks, and minimizes the corresponding waste of target mass. Moreover, whenever the electronic detector reconstruction is compatible with two or more "candidate" bricks, these are ordered by probability and their CS are scrutinized accordingly. This significantly increases the event finding efficiency.

If one or more "event related" tracks are found, the selected brick is exposed to cosmic-rays for about 12 h, thus providing a set of tracks to be used for precise correction of local deformations as required for precision topological and kinematical measurements. The brick is then disassembled and its films are developed. The tracks measured in the CS analysis provide predictions for the so-called scan-back procedure. The latter consists of following a predicted track upstream in the ECC brick until it "disappear". This procedure is initiated in the most downstream film of the brick.

The disappearance of a scan-back track indicates a possible neutrino interaction vertex. A wide area scan is performed over a volume of about 1 cm^3 around the track stopping point, looking for partner tracks and/or secondary decays with a dedicated decay search procedure [103]. This procedure, developed for the tau neutrino search, was successfully applied to the search for charmed hadron production induced by neutrinos. The latter process was indeed used as a control sample to check the efficiency for the detection of the tau lepton, given the similar lifetime of charmed hadrons (about 10^{-12} s). The application of this procedure to muon neutrino interactions led to the observation of 50 decay candidates [103], in good agreement with the expected charmed hadron yield (54 ± 4), derived from the value measured by the CHORUS experiment [104]. Good agreement was found also in the shape of the relevant kinematical and topological variables, like the angle in the transverse plane between the charmed hadron and the muon and the impact parameter of the decay daughter particles with respect to the primary neutrino interaction vertex [103].

Unlike the experiments using "bulk" emulsions like CHORUS where the visual inspection of the primary and decay vertices allows rejecting most of the residual background, the ECC structure prevents the direct check of the vertices for the majority of the events. However, one can still exploit precise kinematical measurements for background suppression. For interesting event topologies, in fact, a detailed kinematical analysis is performed in OPERA by means of the electromagnetic shower energy measurement in the downstream part of the brick, the determination of the momentum by Multiple Coulomb Scattering measurement in the lead/emulsion structure [105], and the connection of tracks in consecutive target walls.

During the five CNGS production runs from 2008 to 2012, OPERA collected about 1.8×10^{20} protons on target and more than 19,000 neutrino interactions. The first tau neutrino candidate was reported in 2010 [106] and the display of its event

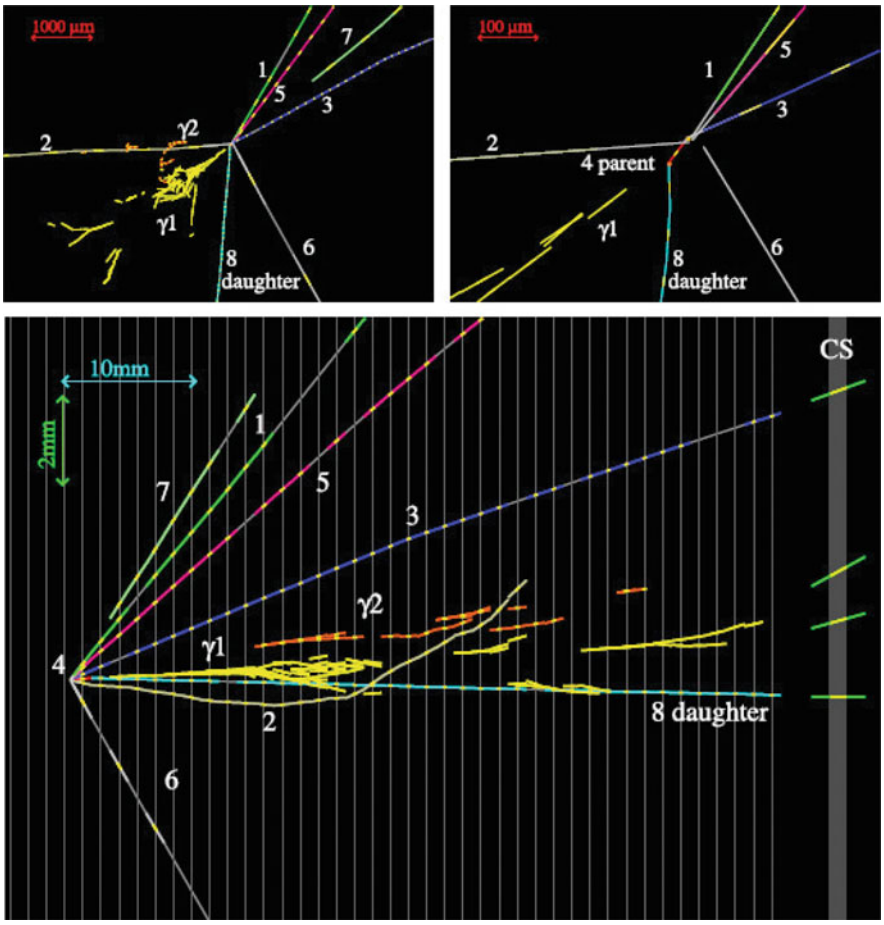


Fig. 9.12 Display of the first ν_τ candidate. Top left: view transverse to the neutrino direction. Top right: same view zoomed on the primary and secondary vertices. Bottom: longitudinal view. Track 4 exhibits a kink topology with an angle of (41 ± 2) mrad after a path length of (1335 ± 35) μm and produces track 8 and the two γ's. Track 2 is identified as a proton and the other charged particles are all consistent with being hadrons [106]

reconstruction is shown in Fig. 9.12. The primary neutrino vertex consists of 7 tracks of which one shows a kink decay topology. None of the primary particles is consistent with neither a muon nor an electron. Two electromagnetic showers induced by γ conversions are visible in Fig. 9.12, indicated as γ_1 and γ_2. These two γs originate from the secondary vertex and their invariant mass is consistent with that of a π^0. From the kinematical analysis performed, the observed decay is consistent with the $\tau \to \rho\nu_\tau$ channel (B.R. $\simeq 25\%$), followed by $\rho \to \pi^0\pi$.

The second [107] and third [108] tau neutrino candidates were reported in 2013, respectively in the $\tau \to \pi\pi\pi\nu_\tau$ and $\tau \to \mu\bar{\nu}_\mu\nu_\tau$ decay channels. The forth

Table 9.1 Overall number of located neutrino interactions with the decay search procedure applied

	2008	2009	2010	2011	2012	Total
p.o.t. (10^{19})	1.7	3.5	4.1	4.8	3.9	18.0
0μ events	150	255	278	291	223	1197
1μ events (p_μ <15 GeV/c)	543	1024	1001	1031	807	4406
Total events	693	1279	1279	1322	1030	5603

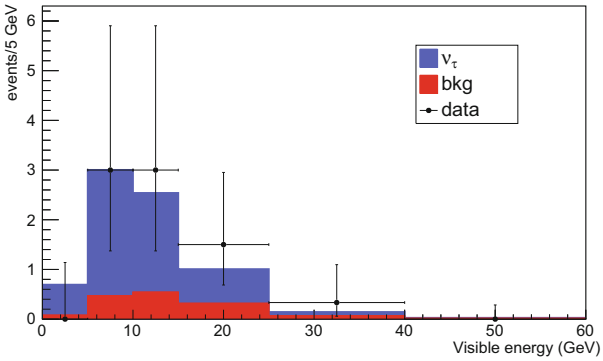


Fig. 9.13 Visible energy distribution of the 10 tau neutrino candidates found in the final sample [114]

candidate was reported in 2014 [109] while the discovery of ν_τ appearance was achieved in 2015 with the observation of a fifth tau neutrino candidate over an expected background of 0.25 events [110]. The OPERA discovery of tau neutrino appearance was explicitly mentioned in the Scientific Background of the 2015 Nobel Prize in Physics.

The emulsion handling was completed in 2015 while the emulsion film scanning was completed in 2016 when the detector was decommissioned. The final number of events passing all the analysis chain up to the decay search are shown in Table 9.1. Events are divided in two categories according to the presence (1μ) or absence (0μ) of a muon in the final state and undergo different selections: a momentum cut of 15 GeV/c is applied to the muons in order to reduce the background.

Given the data-driven validation of the simulation in all corners of the parameter space [103, 111, 112], the OPERA Collaboration decided to release the cuts and exploit the kinematical features of the events with a likelihood approach: this approach enlarges the selected sample, thus reducing the statistical uncertainty for the estimate of the oscillation parameters. Ten tau neutrino candidates were found with the new analysis strategy in the final sample. The distribution of the visible energy for the 10 candidates is shown in Fig. 9.13 together with the expected spectrum.

The number of expected tau neutrino events with looser cuts applied is reported in Table 9.2, together with the number of observed ν_τ candidates in each tau

Table 9.2 Expected signal and background events for the analysed data sample

	Expected background					
Channel	Charm	Had. re-interaction	Large μ-scat.	Total	ν_τ expected	Observed
$\tau \to 1h$	0.15 ± 0.03	1.28 ± 0.38	–	1.43 ± 0.39	2.96 ± 0.59	6
$\tau \to 3h$	0.44 ± 0.09	0.09 ± 0.03	–	0.52 ± 0.09	1.83 ± 0.37	3
$\tau \to \mu$	0.008 ± 0.002	–	0.016 ± 0.008	0.024 ± 0.008	1.15 ± 0.23	1
$\tau \to e$	0.035 ± 0.007	–	–	0.035 ± 0.007	0.84 ± 0.17	0
Total	0.63 ± 0.10	1.37 ± 0.38	0.016 ± 0.008	2.0 ± 0.4	6.8 ± 1.4	10

decay channel. The reported values assume $\Delta m^2_{23} = 2.50 \times 10^{-3}\,\mathrm{eV}^2$ [113] and $\sin^2 2\theta_{23} = 1$. The discovery of tau neutrino appearance is confirmed with a significance of 6.1σ, evaluated by accounting for the features of the events with a likelihood analysis. The increased statistical sample was used to provide the first measurement of Δm^2_{23} in appearance mode with an improved accuracy, giving $\Delta m^2_{23} = (2.7^{+0.7}_{-0.6}) \times 10^{-3}\,\mathrm{eV}^2$ [114].

OPERA has demonstrated the capability of identifying all three neutrino flavours. Emulsion cloud chambers can clearly distinguish between electrons and γs, given their micrometric accuracy emphasizing the displacement between the γ production and conversion vertices. Unlike other detectors, this feature makes the e/π^0 separation particularly efficient and their selection pure: this translates into a very good separation between ν_e charged-current interactions and ν_μ neutral-current ones with a π^0 in the final state. OPERA has searched for the sub-dominant $\nu_\mu \to \nu_e$ oscillations also to constraint the existence of sterile neutrinos. In the analysis of the 2008 and 2009 run data, 19 electron neutrino candidates were found and the results are summarised in [115]. The analysis of the final sample has collected 35 ν_e candidates and the constraints to sterile neutrinos are reported in [116]. Constraints to sterile neutrinos were set also with the analysis of $\nu_\mu \to \nu_\tau$ oscillations [117].

9.6 Future Experiments and Applications

After more than 100 years since its first use, nuclear emulsions are still attractive in a wide range of scientific fields and applications. As it was the case for past developments, the future of nuclear emulsions will again rely on the parallel progress of high-performance readout systems as well as of innovative detector design. We discuss here the cutting edge technology and also review ongoing and emerging applications.

9.6.1 The State-of-the-Art Emulsion Technology

9.6.1.1 High-Performance Scanning Systems

Improvements of scanning systems in speed and quality are continuously progressing. One of the recent breakthrough was the appearance of GPGPU (General Purpose Graphic Processing Unit, or simply GPU). Up to the systems for OPERA, either FPGAs or CPUs were employed for image processing and track reconstruction. The FPGA has a big computing power, but also difficulties in implementing sophisticated algorithms and in flexibility. The CPU can process complicated algorithms but is limited in computing capability. On the other hand, the GPU provides both computing power and flexibility.

The effort to implement GPUs for scanning systems started soon after the release of CUDA [118], and it has quickly become the "standard" in the scanning system development nowadays. The early works were aiming at improving the angular acceptance in track reconstruction which was limited by the lack of computing power for online processing. The previously mentioned S-UTS, the scanning system for OPERA, could recognize tracks with their angle within 30° with respect to the normal of the film surface. This angular acceptance is equivalent to 14% of the entire solid angle. An extension of the S-UTS algorithm was translated into the GPU code, which reconstructed tracks up to 72° (68%) with a reasonable processing time [119]. In parallel, new algorithms suitable for parallel processing were developed to extend the track reconstruction to almost the entire 4π solid angle [120, 121], which finally allowed to fully exploit the 3D tracking capability of nuclear emulsion. Examples of applications of such systems will be discussed further below.

In the data acquisition, there are two complementary ways for the fast readout of emulsion data: maximize the field of view or minimize the dead time due to microscope stage movement. An extreme case of the first approach was implemented in the HTS system (Hyper Track Selector, [122]) as shown in Fig. 9.14, which is the

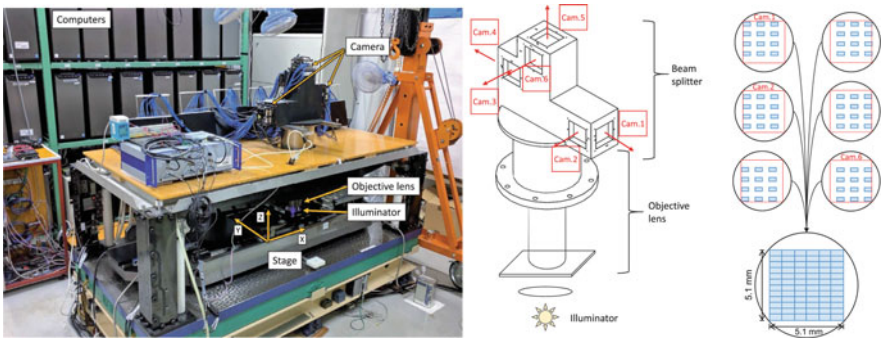


Fig. 9.14 Left: the fast emulsion readout system, Hyper Track Selector (HTS) [122]. Right: the optics and camera system for HTS. The optical path is divided into six mosaic camera modules. Each camera module consists of 12 2.2-Mpixel image sensors. In total 72 image sensors work in parallel to realize a large FOV of 5.1 mm × 5.1 mm with sub-micrometric resolution

fastest readout system at present. Conventional systems were using a field of view (FOV) of 0.12 mm × 0.12 mm (S-UTS) or 0.3 mm × 0.4 mm (ESS). HTS makes use of the custom made objective lens with a large FOV of 5.1 mm × 5.1 mm and a magnification of 12.1. The optical path is divided into six, correspondingly the image is projected on six "mosaic camera modules" as also schematically drawn in Fig. 9.14. Each mosaic camera module consists of 12 2.2-Mpixel image sensors. In total, 72 image sensors work in parallel to build the large FOV. The raw image data throughput from 72 image sensors amounts to 48 GBytes/s, which is then processed in real-time by 36 tracking computers with two GPUs each. The scanning speed has reached 4700 cm^2/h, which is clearly a big leap from the previous generations as shown in Fig. 9.15.

Another approach is to remove the dead time due to the microscope stage movement. In conventional systems, the data taking sequence is the so-called "stop-and-go" where the need to dump stage vibrations limits the repetition cycle up to 6 Hz. In order to minimize this effect, it was proposed to use tomographic image data taking without stopping the stage. In fact, S-UTS was the first system to implement the continuous motion as above mentioned. However, the camera resolution was relatively small (512 × 512 pixels) when compared to the market standard of today. The New Generation Scanning System (NGSS) was developed with a larger camera resolution (2336 × 1728 pixels) and with a different style of continuous motion that allowed running on normal motion hardware of ESS. The schematic of image taking sequence is shown in Fig. 9.16. By realizing a 12 Hz data taking, the scanning speed reached 190 cm^2/h [123].

The advances in scanning speed allows physicists to design experiments with a detector areas of 1000 m^2 to be analysed in a year, to be compared to the total scanned area of OPERA of about 500 m^2 in 5 years. The environment of emulsion readout is continuously changing as long as technologies grow. A new design of scanning system, so called HTS2, is going to combine the large field of view of HTS and the continuous motion [122], which might reach a scanning speed of 5 m^2/h in early 2020s. At this stage, the scanning speed would be no longer a bottleneck of

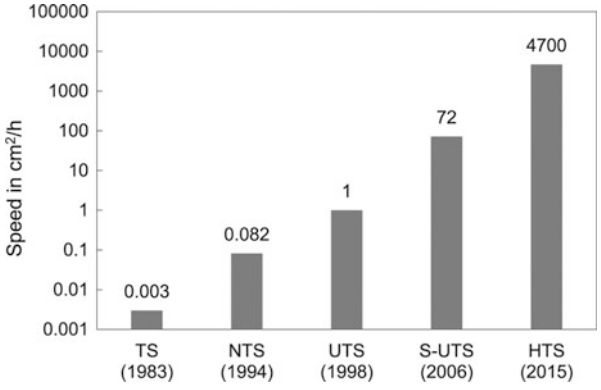


Fig. 9.15 Time evolution of the scanning speed of the Track Selector system. The scanning speed progress in log scale

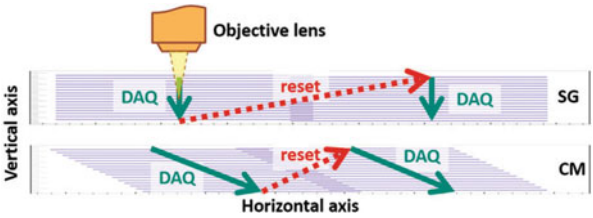


Fig. 9.16 Schematic drawing of the Stop and Go (SG) motion and Continuous Motion (CM) of NGSS [123]

any experiment and new challenging experiments might be proposed, based on such a high-speed readout framework.

9.6.1.2 Fine-Grained Emulsion Production

Owing to its unbeatable position and angular resolution, the emulsion technique is being adopted in different applications in the fields of fundamental physics and applied science. The OPERA film [2], which was mass produced in industries, has been used for some applications, although the properties of the detector are tuned for the OPERA experiment. Following the increased interest in using emulsion detectors in a broad range of applications, the R&D of emulsion gel has become essential for optimising the detector for each application. However, conducting R&D for each small-scale experiment is difficult in industrial companies. This motivated the Nagoya University group to set up their own emulsion gel production facility in 2010. With the help of experts from FUJI Film Co. Japan, custom-made emulsion gels were successfully produced with an improved sensitivity to minimum ionizing particles with respect to OPERA films [3]. Moreover, some R&D programs were conducted to control silver halide crystal size, which defines spatial resolution and sensitivity. Fine-grained emulsions were produced with a crystal size of a few tens of nanometres, which is approximately one order of magnitude smaller than the conventional one (Fig. 9.17). They are called Nano Imaging Trackers (NIT) [124, 125]. The average size of NIT crystals was measured to be 44.2 ± 0.2 nm, with a standard deviation of 6.8 nm. NITs are not sensitive to the minimum ionizing particles but are sensitive enough to low-velocity heavy ions. They are considered a possible detector for detecting the recoiled nuclei induced by dark matter.

9.6.1.3 Large Grain Emulsion Production

For certain applications such as muon radiography, large-scale detectors are required. An improvement in the readout speed is therefore crucial to make future large-scale applications possible, and the availability of a new type of emulsion featuring crystals of larger sizes is one way to pursue this goal. This would allow a lower magnification for the microscopes and, consequently, a larger field of view resulting in a faster data analysis. The size of the crystals used for the neutrino

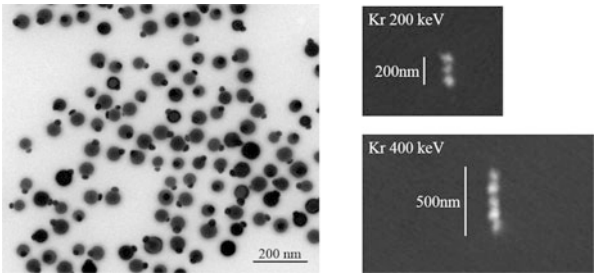


Fig. 9.17 Left: Silver halide crystals in the fine-grained emulsion [124, 125], as seen with a transmission electron microscope. Photolytic silver grains are also visible on the surfaces of silver halide crystals. Right: Tracks of Kr ions in such an emulsion, as seen with a scanning electron microscope

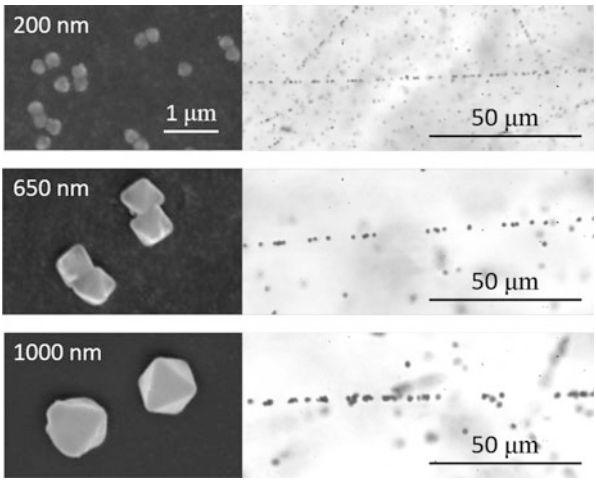


Fig. 9.18 Electron microscope pictures of silver halide crystals (left) and electron tracks (right) in a conventional film and in the newly developed samples [126]

oscillation experiments mentioned above was 200 nm and has never been larger than 300 nm in previous experiments. The production of new types of emulsions with crystal sizes of 600–1000 nm, 3–5 times larger than those of standard films, has been studied and realised using the gel production machine at Nagoya University. The first results characterising newly produced emulsions have been reported [126], showing a sufficient sensitivity and a good signal to noise ratio (Fig. 9.18). This development will allow a 25 times faster readout speed by using lower magnification objective lenses. These new detectors will pave the way to future large-scale applications of the technology, e.g. 3D imaging using muon radiography or future neutrino experiments.

In close connection with the production of large crystals, there has also been a study to produce crystals slightly larger (350–400 nm) than 200 nm and to check

the dependence of crystal sensitivity on its size. This study was motivated by the interest to understand the phenomenology of the latent image formation, which predicts that the quantum sensitivity can be better at such a crystal size. Further studies are in progress with the aim of developing emulsions with higher sensitivity. The conditions of chemical sensitisation and development were optimised for each crystal size in the range of 200–800 nm. The increase in the crystal sensitivity depending on the crystal size was confirmed for crystals of 350–800 nm [127]. These R&D activities form the base for a broad range of future applications.

9.6.2 Projects in Fundamental Physics

9.6.2.1 Balloon Experiments

Balloon experiments employing emulsion detectors were reported in [128]. The use of emulsion technique for cosmic-ray physics experiments has recently attracted research interest after the significant technological advances in the last decades. In 2004, a balloon experiment using emulsions was performed to observe primary cosmic-ray electrons [129]. Various innovations such as the industrial emulsion films, the refreshing technique, the automated emulsion read-out system and the off-line analysis methods were introduced. In addition, a dedicated device was developed to distinguish between particles passing through a chamber at the balloon level altitude and those recorded during other periods. The mechanism of this device is such that it causes intentional shift of the upper block of the chamber with respect to the lower block, when the balloon reaches float altitude and again when the flight at float altitude is terminated. The working principle of the technique was successfully demonstrated.

Based on these techniques, a balloon-borne emulsion γ-ray telescope was proposed [130] and the Gamma-ray Astro-Imager with Nuclear Emulsion (GRAINE) project was developed for the observation of γ-rays in the energy range of 10 MeV–100 GeV. A precise, polarisation-sensitive, large-aperture-area emulsion telescope with repetitive long-duration balloon flights was employed. The electron and positron angles at the pair creation point can be measured in emulsions and the angular resolution for γ-rays (10 MeV–10 GeV) is about one order of magnitude higher than that of the Fermi Large Area Telescope (Fermi-LAT) (Fig. 9.19). The polarisation sensitivity of an emulsion-based telescope was demonstrated using a polarised γ-ray beam at SPring-8/LEPS [131].

An emulsion multi-stage shifter was used to develop an innovative solution capable of providing the event time-stamp and hence the γ-ray absolute direction [135]. The relative alignment between the automatically sliding emulsion films, each moving with a known different speed, provides the required time association of the event. This technique allows γ-ray detection with low energy threshold, minimising the electric power and also limiting the overall detector mass.

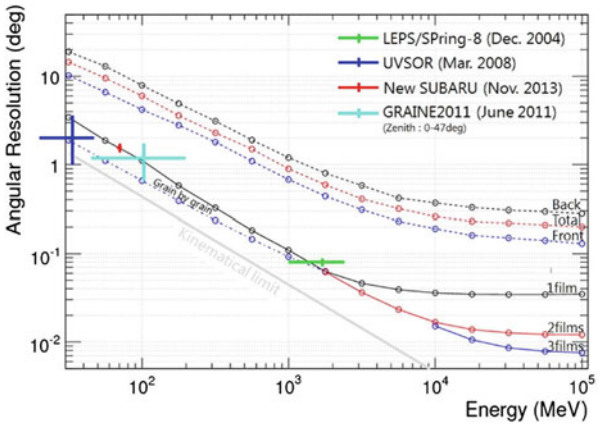


Fig. 9.19 Angular resolution of the emulsion γ-ray telescope (lines show simulation results and dots with error bars show experimental results) [132]. The measurements were performed with γ-ray beams (LEP/SPring-8, UVSOR, and New SUBARU) and using a flight data [133]. Dotted lines show angular resolution with Fermi-LAT for the front section with thin radiation foils and the back section with thick foils [134]

In 2011, the first balloon-borne experiment was performed with a $12.5 \times 10\,\mathrm{cm}^2$ aperture area and 4.6-h flight duration for a feasibility test [133]. The chamber comprised three sections. The top section was made of an ECC with emulsion films interleaved with copper foils (50 μm), meant to measure the γ-ray angle around the conversion point. The middle section included an emulsion multi-stage shifter, providing the time-stamp of the events. The bottom part contained a calorimeter comprising a lead/emulsion ECC for the γ energy measurement. With this flight data, systematic detection, energy reconstruction, and timestamping of γ-ray events were performed [133] and subsecond time resolution of the emulsion γ-ray telescope was demonstrated [136]. The second balloon-borne experiment was performed at the Alice Springs balloon-launching station in 2015 [137]. The telescope had a $3780\,\mathrm{cm}^2$ aperture and was taking data for a total of 14.4 h. The experiment aimed at demonstrating the overall performance of the emulsion γ-ray telescope. The improvements in the emulsion characteristics and handling applied to this experiment are summarised in [138]. The project plans a third balloon-borne experiment in 2018 for the celestial source detection and envisions scientific observations from 2021.

9.6.2.2 The NEWSdm Experiment

The nature of Dark Matter is one of the fundamental questions to be answered. Direct Dark Matter searches are focussed on the development, construction, and operation of detectors looking for the scattering of Weakly Interactive Massive

Particles (WIMPs) with target nuclei. The measurement of the direction of WIMP-induced nuclear recoils is a challenging strategy to extend the sensitivity of dark matter searches beyond the neutrino-induced background event rate and provide an unambiguous signature of the detection of Galactic dark matter [139]. Current directional experiments are based on the use of gas TPC whose sensitivity is strongly limited by the small achievable detector mass. Nuclear Emulsions for WIMP Search with directional measurement, NEWSdm, is an innovative directional experiment proposal based on the use of a solid target made by newly developed nuclear emulsion films and read-out systems capable to detect nanometric trajectories.

The approach proposed by the NEWSdm Collaboration [140] consists of using a nuclear emulsion-based detector acting both as target and as nanometric tracking device. The NEWSdm project foresees the employment of NIT. The detector is conceived as a bulk of NIT surrounded by a shield to reduce the external background. The detector is then placed on an equatorial telescope in order to absorb the Earth rotation, thus keeping fixed the detector orientation with respect to the incoming apparent WIMP flux. The angular distribution of the WIMP-scattered nuclei is therefore expected to be strongly anisotropic with a peak centred in the forward direction.

NIT have a linear density of about 11 crystals/μm [124], thus making the reconstruction of trajectories with path lengths as short as 100 nm possible, if analysed by means of microscopes with enough resolution. The presence in the emulsion gel of lighter nuclei such as carbon, oxygen and nitrogen, in addition to the heavier nuclei of silver and bromine, is a key feature of the NEWSdm project, resulting in a good sensitivity to WIMPs in the mass range between 10 to 100 GeV/c^2.

In the NEWSdm experiment a WIMP signal consists of short-path, anisotropically distributed, nuclear recoils over an isotropically distributed background. The search for signal candidates requires the scanning of the whole emulsion volume. The read-out system has therefore to fulfil two main requirements: a fast, completely automated, scanning system is needed to analyse the target volume over a time scale comparable with the exposure; the spatial resolution achievement has to go well beyond the diffraction limit, in such a way to ensure high efficiency and purity in the selection of signal candidates. The analysis of NIT emulsions is performed with a two-step approach: a fast scanning with a state-of-the-art resolution for the signal pre-selection followed by a pin-point check of preselected candidates with unprecedented nanometric resolution to further enhance the signal to noise ratio.

In the first analysis phase, a fast scanning is performed by means of an improved version of the optical microscope used for the scanning of the OPERA films [141]. An R&D program has achieved a speed of about 200 cm^2/h [123, 142].

The starting point of the emulsion scanning is the image analysis to collect clusters making up silver grains. Given the intrinsic resolution of the optical microscope (∼200 nm), the sequence of several grains making a track of a few hundred nanometers may appear as a single cluster. Nevertheless, a cluster made of several grains tends to have an elliptical shape with the major axis along the direction of the trajectory, while a cluster produced by a single grain tends to have a

spherical shape. The shape analysis with an elliptical fit is indeed the first approach to select signal. In order to simulate the effect of a WIMP-induced nuclear recoil and to measure the efficiency and the resolution of the new optical prototype, a test beam with low velocity ions was performed. Kr ion beams with energies of 200 and 400 keV [143] and C ion beams with energies of 60, 80 and 100 keV were used. Silver grains belonging to the tracks appear as a single cluster. An elliptical fit of the cluster shape allows a clear separation between fog grains and signal tracks [144].

The second analysis step at the microscope makes use of the plasmon resonance effect occurring when nanometric silver grains are dispersed in a dielectric medium [145]. The polarization dependence of the resonance frequencies strongly reflects the shape anisotropy and can be used to infer the presence of non-spherical nanometric silver grains within a cluster made of several grains. NEWSdm is using this technology to retrieve track information beyond the diffraction limit. Images of the same cluster taken with different polarization angles show a displacement of the position of its barycentre. The analysis of this displacement allows to distinguish clusters made of a single grain from those made of two or more grains building up a track, as shown in Fig. 9.20: unlike the single grain reported in the top

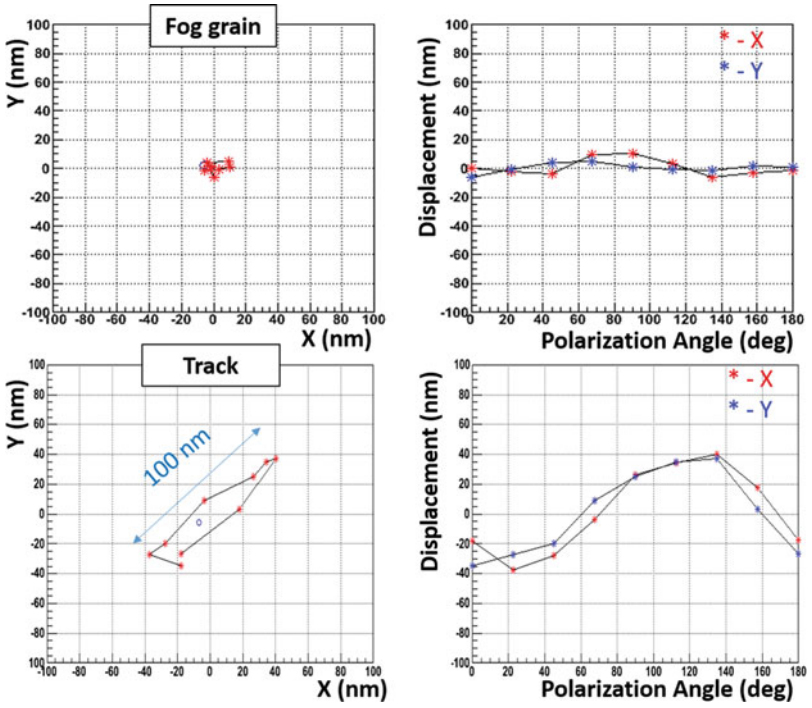


Fig. 9.20 Displacement of the barycentre as a function of the light polarization angle. The response to a single grain (top) and to a C ion track (bottom) are compared. The ion track shows a clear displacement

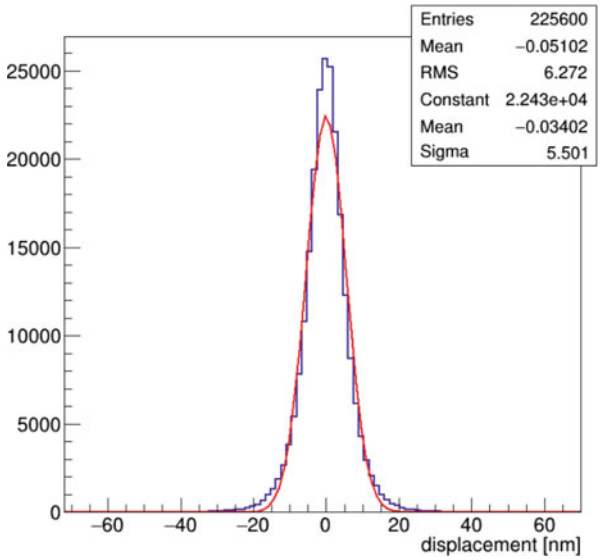


Fig. 9.21 Barycentre displacement of clusters made by single grains

plots, the Carbon ion track in the bottom plot shows a barycentre displacement of 100 nm length while changing the polarization angle. An unprecedented nanometric accuracy has been achieved in both coordinates with this method: Fig. 9.21 reports the displacement of the barycentre of clusters made of single grains, showing an RMS smaller than 10 nm. Such an achievement allows to detect path lengths where the barycentre displacement induced by the polarization change is only a few tens of nanometres. The actual threshold achievable on path lengths depends on the crystal size and can in principle be reduced to a few tens of nanometres as well.

The wavelength of the scattered light depends on the size of the grains where light is scattered off. In order to exploit this effect, the latest version of the optical microscope makes use of a colour camera, thus providing sensitivity to the sense of the track, since grains are expected to be larger at the end of the track range and therefore the scattered light shifts to the red colour. The prototype of this new system in operation in Naples is shown in the left plot of Fig. 9.22 while the image of an α track is reported on the right: different grains show different colours due to their different size, that in turn can provide sensitivity to the particle sense.

The NEWSdm collaboration has installed at the Gran Sasso underground laboratory a facility for the emulsion handling and film production. Moreover, a dedicated structure was constructed in the Hall B of the underground Gran Sasso Laboratory early in 2017 to shield a detector of 10 g mass against the environmental background sources over an exposure time of about 1 month. The experimental setup consists of a shield from environmental backgrounds, made of a few tons of polyethylene and lead, and a cooling system to ensure the required temperature level to the NIT emulsion detector. The aim is to measure the detectable background from environmental and intrinsic sources and to validate estimates from

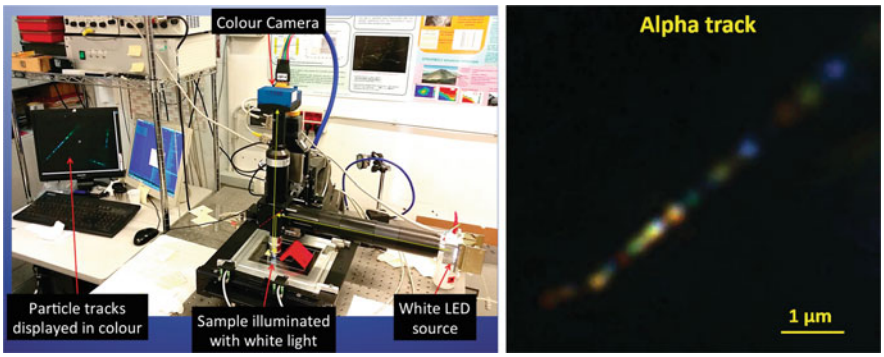


Fig. 9.22 Left: new optical microscope equipped with colour camera in Naples. Right: the last few microns of an α track path showing grains of different colours

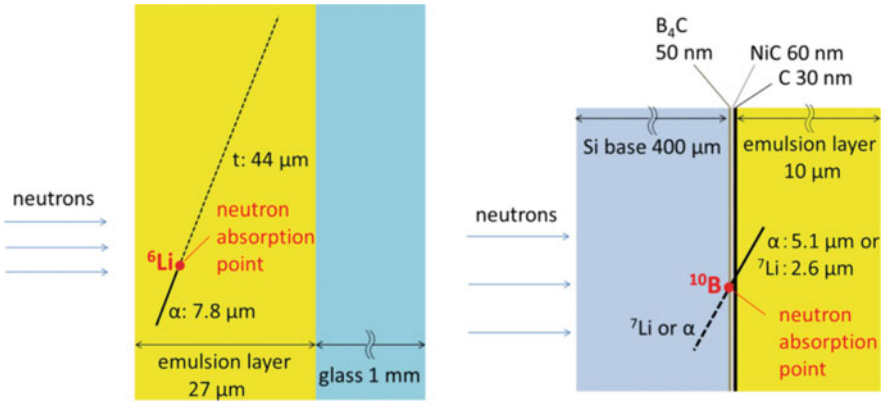


Fig. 9.23 Cross-sectional view of the detectors [149]. Detectors with $LiNO_3$ doping (left) and the $^{10}B_4C$ thin layer (right)

simulations [146]. The confirmation of a negligible background would pave the way for the construction of a pilot experiment with an exposure of about 10 kg year.

9.6.2.3 Development of Cold-Neutron Detector

A new detector for detecting cold and ultra-cold neutrons has been recently developed. It employs fine-grained emulsion detectors with 35-nm-diameter crystals and nuclides with large neutron absorption cross sections, such as ^{6}Li and ^{10}B. One detector type is realised by doping $LiNO_3$ into fine-grained emulsion detectors [147]. The cross-sectional view of the detector is shown in Fig. 9.23 (left). An α particle and a tritium are emitted during the reaction: $^6\text{Li} + \text{n} \rightarrow \alpha + \text{t} + 4.78\,\text{MeV}$. Events of neutron absorption by ^{6}Li were successfully observed by exposing the

detector to thermal neutrons at the Kyoto University Research Reactor Institute (KURRI). The spatial resolution achieved in the measurement of the absorption point was estimated to be 0.34 μm from the average grain density of the track far from the end of its range. The detection efficiency was measured by exposing the detector to a cold neutron beam at BL05 port in the Materials and Life Science Experimental Facility (MLF) at J-PARC [148]. The measured efficiency of $(3.3 \pm 0.6) \times 10^{-4}$ was consistent with the expectation.

The other detector type consists of a 50-nm-thick converter layer made of $^{10}B_4C$ formed on a 0.4-mm-thick silicon substrate and coated by 10-μm-thick, fine-grained emulsion [149]. The converter layer was covered by C (50 nm) and NiC (60 nm) layers. An α particle or a ^{7}Li nucleus will be detected in the emulsion, as shown in Fig. 9.23 (right). They are produced via the reactions: $^{10}B + n \rightarrow \alpha + {}^7Li + 2.79\,MeV$ (6%) or $^{10}B + n \rightarrow \alpha + {}^7Li + 2.31\,MeV$ (94%). The detector was exposed to cold and ultra-cold neutrons at J-PARC, and the events of neutron absorption by ^{10}B were clearly observed. The position resolution of the absorption point in the $^{10}B_4C$ layer depends on the track angle. By limiting the track angle, the expected position resolution is ∼100 nm [150], which is 1–2 orders of magnitude higher than that of the conventional detectors used for detecting cold and ultra-cold neutrons. Further optimisation of the thickness of the converter layer and development of automatic track reconstruction are explored. The development of these detectors paves the way to future applications such as the precise measurement of the position distribution of quantised states of ultra-cold neutrons or neutron imaging with future neutron sources.

9.6.2.4 Study of Double-Hypernuclei

The knowledge of $\Lambda - \Lambda$ interaction is limited as only one out of the nine double hypernuclei events detected by E373 is fully analysable to extract information of the interaction (Fig. 9.24). In order to answer questions such as the nuclear mass dependence of $\Lambda - \Lambda$ interaction, the E07 experiment at J-PARC is being carried out, which is aiming at studying $\Lambda - \Lambda$ interactions with 100 double hypernuclei events, one order of magnitude larger statistics with respect to E373. As schematically shown in the right side of the Fig. 9.24, E07 uses a 1.7 GeV/c K^- beam hitting a diamond target to produce Ξ^- hyperons (dss), subsequently stopped and captured by one of emulsion detector nuclei to produce double hypernuclei. The detector has a hybrid structure with a silicon strip detector and a spectrometer system for the K^+ identification, needed to tag Ξ^- hyperons. The emulsion detector will be analysed by an automated scanning system dedicated to this experiment. E07 conducted the physics runs in 2016 and 2017. The emulsion readout and analysis are in progress.

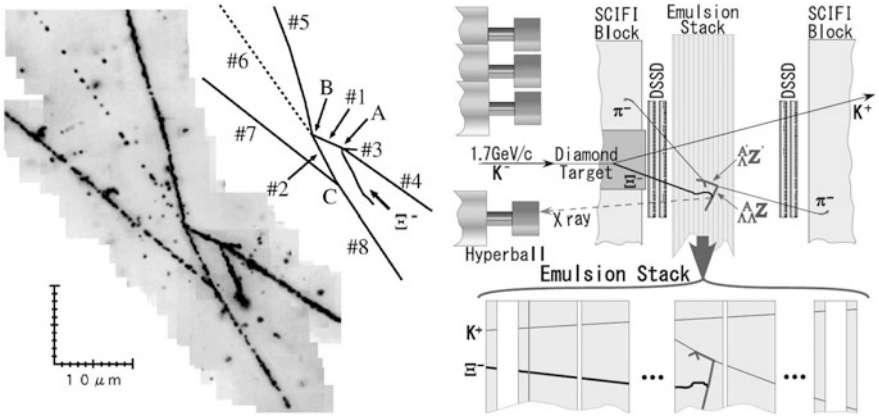


Fig. 9.24 Left: The so-called "Nagara" double hypernucleus event found in the E373 experiment at KEK. Ξ^- was captured at rest by a carbon nucleus in the emulsion detector and produced a double hypernucleus ($^{6}_{\Lambda\Lambda}$He), which decayed in series, leaving a peculiar event topology in emulsion [60]. Right: A schematic of the experimental setup of the E07 experiment at J-PARC [151]

9.6.2.5 Measurements of Antimatter

Emulsion detectors have been recently considered as high-accuracy position sensitive detectors for low-energy antimatter studies. These studies include the AEgIS experiment at CERN [152, 153], with the goal of measuring the Earth's gravitational acceleration on antihydrogen atoms to the ultimate precision of 1%. The vertical deflection of the $\bar{H}$ atoms due to gravity will be detected by a setup comprising material gratings coupled with a position-sensitive detector. The position detector requires the best possible position resolution, which currently is provided by emulsion-based detectors.

There were technical challenges for emulsion detectors to be operated in vacuum and at cryogenic temperatures. In vacuum, water loss leads to cracks in the emulsion layer and an increase in random noise due to mechanical stress caused by the drying process. Two ways of solving these problems were then established. One is to mix glycerin with emulsion gel to replace water with glycerin, and the other one is to put gas barrier films on emulsions to keep water in the films. Both approaches have proven to work in ordinary vacuum (10^{-7}–10^{-5} mbar). At 77 K, the performance of emulsion detectors was not well known. The sensitivity of the emulsion at 77 K was studied and observed to be 43% of the value at 300 K. By optimising the track reconstruction, detecting minimum ionizing particles with such a sensitivity will be feasible since the tracking efficiency for tracks with more than 10 grains in an 50–100 μm thick emulsion layer could be close to 100%. In 2012, an antiproton exposure to the emulsion detectors was performed for the feasibility study at the Antiproton Decelerator (AD [154]) at CERN. Fig. 9.25 shows an annihilation vertex on the bare emulsion surface. Annihilation vertices in the metal target were

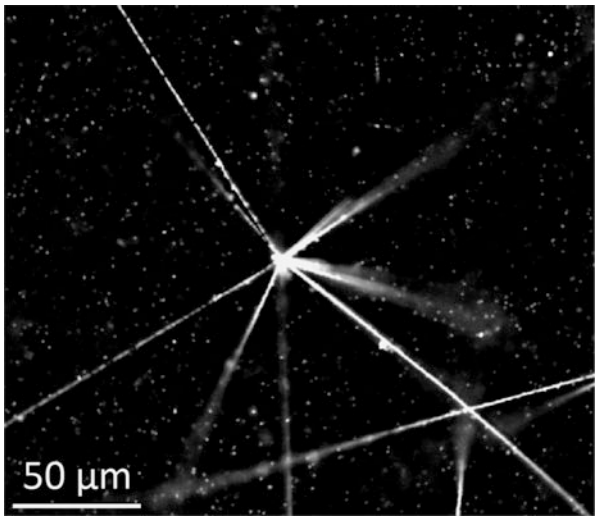


Fig. 9.25 An antiproton annihilation vertex in an emulsion layer [155]. The view is perpendicular to the antiproton beam direction

also reconstructed, demonstrating a resolution of 1 µm on the vertical position. In addition, a proof-of-principle experiment with mini-moiré deflectometer was performed [156]. The periodic patterns were observed as expected in the emulsions, and the measured shift between antiprotons and light was consistent with the force from the magnetic field at the given position. The results are a crucial step toward the direct detection of the gravitational acceleration of antihydrogen. In 2014, measurements of the multiplicities of charged annihilation products on different target materials, namely copper, silver, and gold, were performed [157] at the CERN AD. Apart from the obvious applications in nuclear physics, this measurement can provide a useful check of the ability of standard Monte Carlo packages to reproduce fragment multiplicities and energy distributions. The measured fragment multiplicities were not well reproduced by the different models used in Monte Carlo simulation with the exception of FLUKA [158, 159], which is in good agreement with the particle multiplicities for both minimum and heavily ionizing particles (Fig. 9.26).

There is another proposal by the QUantum interferometry and gravity with Positrons and LASers (QUPLAS) project to use emulsions for their studies on positrons [160]. The sensitivity of the emulsion detectors was studied using a mono-energetic positron beam at energies as low as 9–18 keV. The obtained results prove that the emulsions are highly efficient at detecting positrons at these energies. This achievement paves the way to perform matter-wave interferometry with positrons using this technology.

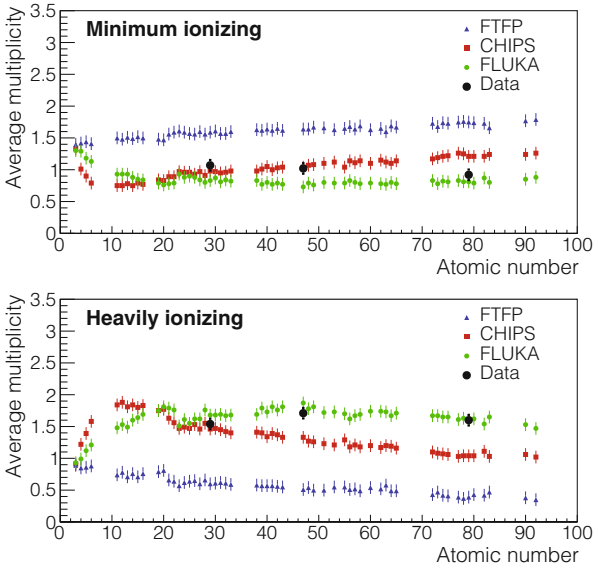


Fig. 9.26 Particle multiplicity from antiproton annihilations as a function of atomic number for minimum and heavily ionizing particles [157]

9.6.2.6 Accelerator Beam Characterization: Muon Measurements at the T2K ν Beamline

The high spacial resolution of emulsion detectors turns out to have an advantage in the characterization of high-intensity accelerator beams, in particular in fast extraction mode where billions of particles arrive within a nanosecond: in these conditions electronic detector cannot identify particles on an event by event base given their limited occupancy. A notable example is the muon measurement at the T2K neutrino beam from J-PARC in Japan [161]. As neutrinos and muons are both produced by meson decays ($\pi, K \rightarrow \mu\nu_\mu$), the understanding of the muons provides valuable information about neutrinos, such as the parent hadron production and momentum distribution. Nevertheless, low energy electromagnetic components highly contaminate the muon flux at the muon pit downstream of the decay volume ($\mu^\pm = 53\%$, $e^\pm = 7\%$, $\gamma = 40\%$, estimated by MC). This makes it difficult to extract meaningful information from the muon beam. The muons are regularly measured by the silicon photodiodes and ionization chambers at the muon pit to monitor the beam direction in each spill. These are charge-integration detectors not optimised to measure muon tracks. Therefore, a measurement of muons by means of the emulsion detectors was performed. The emulsion detector module was composed of 8 OPERA-type films with a cut-off momentum of 30 MeV/c for electrons given by the dedicated track recognition procedure, efficiently achieving a 99% purity of muons after reconstruction. The

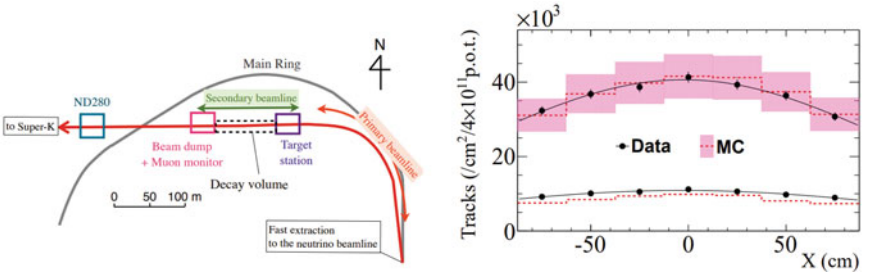


Fig. 9.27 Left: A schematic of the T2K neutrino beamline. The muon measurement was performed at the muon monitor pit behind the decay volume. Right: Comparison of the muon flux with the prediction at the horn current of 250 kA (top) and off (bottom). Figures from [161]

measurement was done at an intensity of the order of 10^{11} protons on target, which yielded $O(10^4)$ muons/cm^2 in the emulsion detectors. The measured profile is shown in Fig. 9.27 with the expected profile predicted by the FLUKA simulation with dedicated hadron production tuning. The absolute muon flux was measured for the first time at neutrino beamlines, thus characterising the beamline. In addition to the flux measurement, the momentum distribution of muons has been measured with the OPERA-like ECC with 25 films sandwiched with 24 1-mm-thick lead plates, which is to be published.

Such muon measurements can be performed at future neutrino beamlines e.g. the J-PARC neutrino beamline for the Hyper-K experiment [162] and the LBNF (Long-Baseline Neutrino Facility) for the DUNE experiment [163]. In general, nuclear emulsions provide unique capability to study high-intensity accelerator beamline operated in fast extraction mode. Thanks to the automated scanning system, this field is expected to grow in the future.

9.6.2.7 The NINJA Project

The Neutrino Interaction research with Nuclear emulsion and J-PARC Accelerator (NINJA) project was initiated for the precise measurement of neutrino–nucleus interactions. The study of neutrino–nucleus interactions in the sub-multi-GeV region is important to reduce systematic uncertainties in present and future neutrino oscillation experiments. The emulsion detector can measure particles with a low energy threshold for various targets such as iron, carbon and water. It also exhibits good electron/gamma separation capability, allowing for precise measurements of electron–neutrino interactions. Given these capabilities, the future program includes searches for sterile neutrinos with a detector made of three components. The upstream part is made of an ECC with emulsion films interleaved by the target material, which is used for detecting neutrino interactions. The middle part includes an emulsion multi-stage shifter device [164], providing the timing information of

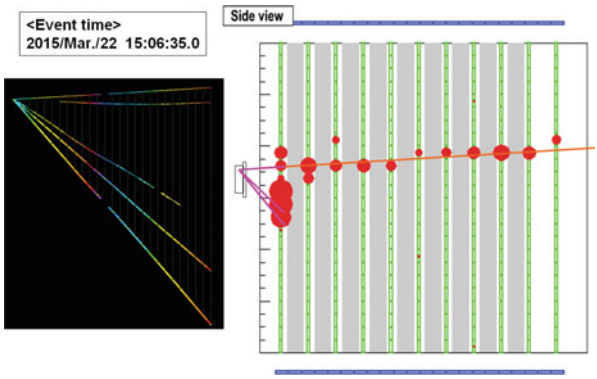


Fig. 9.28 Hybrid analysis of ECC and INGRID (side view of an event) [166]

the events. The downstream part is the Interactive Neutrino Grid (INGRID), which is one of the near detectors for the T2K experiment [165] used to identify muons.

A test experiment (J-PARC T60) was implemented as a first step in the project to check the performance of the detector. Its neutrino event analysis is based on scanning the full area of the emulsion detectors by the HTS system. A more detailed analysis of the detected events could be performed by a dedicated scanning procedure with an extended angular acceptance [119]. The feasibility of the project was studied in the first exposure to a 2-kg iron target in 2015. The full area of the emulsion films ($\sim$1.2 m^2) was scanned and a systematic analysis was performed to locate neutrino interactions. The neutrino candidate events located in the emulsions were matched to events observed by INGRID by employing the timing information from the multi-stage shifter. The hybrid analysis of ECC and INGRID was demonstrated [166]. An analysed event is shown in Fig. 9.28. Further, some other exposures of the anti-neutrino beam to the detectors were conducted, testing the first proto-type of water-target ECC or checking the detector performance with higher statistics [167]. The plan is to scale up the detector step-wise. Physics runs will be planned based on the results of these test experiments.

9.6.2.8 Tau-Neutrino Production Studies

At the CERN Super Proton Synchrotron (SPS), a new project called DsTau has been proposed to study tau-neutrino production [168] aiming at providing important information for future ν_τ measurements where high ν_τ statistics is expected. The results of DsTau are a prerequisite for measuring the ν_τ charged-current cross section, which has never been adequately measured (only the DONUT measurement was reported so far [87]). Precise measurement of the cross section would enable a search for new physics effects in ν_τ–nucleon CC interactions. It also has practical implications for neutrino oscillation experiments such as Super-K, Hyper-K [162] and DUNE [163], which suffer from a ν_τ background to their ν_e measurements. As for the DONUT experiment, the dominant source of ν_τ is the sequential decay of D_s

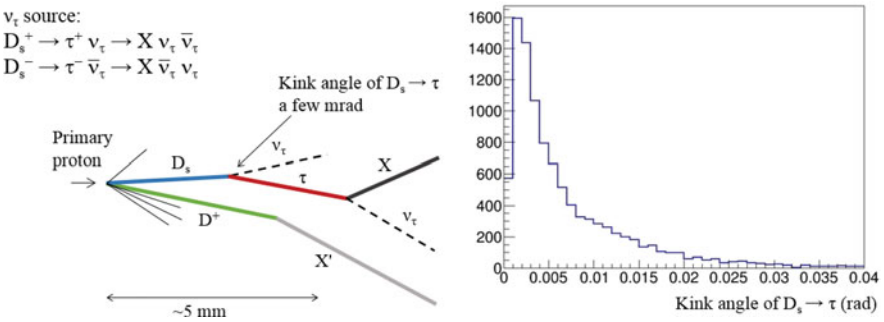


Fig. 9.29 Topology of $D_s \to \tau \to X$ events (left) and simulated kink angle distribution of $D_s \to \tau$ (right) [168]

mesons, $D_s^+ \to \tau^+ \nu_\tau \to X \nu_\tau \overline{\nu}_\tau$ and $D_s^- \to \tau^- \overline{\nu}_\tau \to X \overline{\nu}_\tau \nu_\tau$ produced in high-energy proton interactions. The topology of such an event is shown in Fig. 9.29. Directly measuring $D_s \to \tau$ decays will provide an inclusive measurement of the D_s production rate and the decay branching ratio to τ. The D_s momentum will be reconstructed by combining the topological variables measured in the emulsion detector.

The project aims at detecting 10^3 $D_s \to \tau$ decays to study the differential production cross section of D_s mesons. For this purpose, emulsion detectors with a nanometer precision readout will be used. An emulsion detector with a crystal size of 200 nm has a position resolution of 50 nm [3], as shown in Fig. 9.2, allowing for kink detection with a threshold of 2 mrad at the 4σ confidence level. The global analysis will be based on fast scanning of the full area by the HTS system [122]. After the τ decay trigger, the events will be analysed by dedicated high-precision systems [120] using a piezo-based high-precision z-axis, allowing the emulsion hits to be measured with a nanometric resolution. Each detector unit consists of a 500 μm-thick tungsten target, followed by 10 emulsion films interleaved with 200 μm-thick plastic sheets acting as decay volumes for short-lived particles as well as high-precision particle trackers. Ten such units are used to construct a module, which is followed by an ECC to measure the momenta of the daughter particles. With this module, 4.6×10^9 protons on target are needed to accumulate 2.3×10^8 proton interactions in the tungsten plates. The data generated by this project will enable the ν_τ cross section measured by DONUT to be re-evaluated, which should significantly reduce the total systematic uncertainty. Once ν_τ production is established, the next stage will be to increase the number of ν_τ detected events. This could be achieved within the framework of the SHiP project [171] at CERN because its beamline (beam-dump type) is well suited for this task. The DsTau project aims to look for new physics effects in ν_τ–nucleon CC interactions with a total uncertainty of 10%. In addition to the main aim of measuring D_s, analysing 2.3×10^8 proton interactions, combined with the high yield of 10^5 charmed decays produced as by-products, will enable the extraction of additional physical quantities. Based on the

results of beam tests undertaken in 2016 and 2017 for the feasibility study, a pilot run is scheduled for 2018 and physics runs are planned from 2021 after the upcoming long shutdown of the accelerator complex at CERN.

Another proposal, called the SHiP-charm project [169], aims at measuring the associated charm production by employing the SPS 400 GeV/c proton beam. Charmed hadrons are produced either directly from interactions of the primary protons or from subsequent interactions of the particles produced in the hadronic cascade showers. Recent detailed simulation studies of proton interactions in heavy and thick targets show a sizeable contribution from the cascade production to the charmed hadron yield [170]. This proposal includes a study of the cascade effect to be carried out using ECC techniques, i.e. slabs consisting of a replica of the SHiP experiment target [171] interleaved with emulsion films. The detector is hybrid, combining the emulsion technique with electronic detectors to provide the charge and momentum measurement of charmed hadron decay daughters and the muon identification. This allows a full kinematical reconstruction required by the double-differential cross-section measurement. According to the simulation performed, the delivery of 2×10^7 protons on target would allow the detection of about 1000 fully reconstructed charmed hadron pairs. An optimisation run is scheduled for 2018 and the full measurement is planned after the long shutdown LS2 of the CERN accelerator complex, with 5×10^7 protons on target and a charm yield of about 2500 fully reconstructed interactions.

These two approaches, DsTau and SHiP-charm, are complementary since DsTau will detect 10^5 charmed hadron pairs with good D_s selection capability, while SHiP-charm will study about 2500 fully reconstructed charmed hadron pairs including the hadronic cascade effect. The results of these approaches will provide essential input for future ν_τ measurements.

9.6.2.9 The SHiP Experiment

The discovery of the Higgs boson in 2012 has fully confirmed the Standard Model of particles and fields. Nevertheless, there are still fundamental phenomena, like the existence of dark matter, the baryon asymmetry of the Universe and the origin of neutrino masses, that could be explained by the discovery of new particles. Searches for new physics with accelerators are performed at the LHC, looking for very massive particles coupled to matter with ordinary strength. A new experiment, Search for Hidden Particles (SHiP), has been proposed [171], designed to operate at a beam dump facility to be built at CERN and to search for weakly coupled particles in the few GeV mass range. A beam dump facility using high intensity 400 GeV/c protons would be a copious source of such unknown particles in the GeV mass range. Since a high-intensity tau neutrino flux is produced by such a facility from D_s decays, the experimental apparatus foresees a neutrino detector to study the tau neutrino cross-section and discover the tau anti-neutrino. This detector is also suited to detect dark matter or any weakly interacting particle through its scattering off the

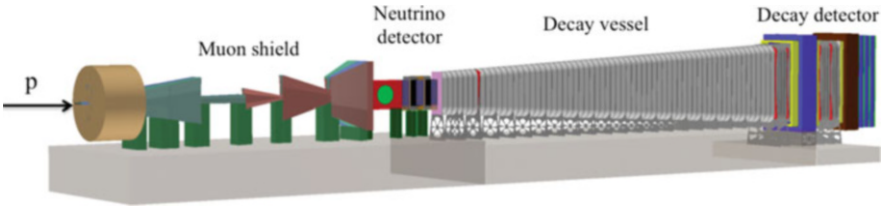


Fig. 9.30 Layout of the SHiP project

atoms of the apparatus target. The physics case for such an experiment is widely discussed in [172].

Figure 9.30 shows the SHiP facility to be placed in the North Area. In 5 years, the facility will integrate 2×10^{20} 400 GeV/c protons, produced by the SPS accelerator complex, impinging on a 12 interaction length (λ_{int}) target made of Molybdenum and Tungsten, followed by a 30 λ_{int} iron hadron absorber. Downstream of the target, the hadron absorber filters out all hadrons, therefore only muons and neutrinos are left. An active muon shield [173] is designed with two sections with opposite polarities to maximize the muon flux reduction: it reduces the muon flux from $\sim10^{10}$ down to $\sim 10^5$ muons per spill. Approximately 4×10^{13} protons are extracted in each spill, designed to be 1 s long to reduce the detector occupancy. The tau neutrino detector is located downstream of the muon shield, followed by the decay vessel and the detector for hidden particles.

The neutrino detector is made of a magnetised target region, followed by a muon spectrometer. The neutrino target is based on the emulsion cloud chamber technology employed by the OPERA experiment, with a compact emulsion spectrometer, made of a sequence of very low density material and emulsion films to measure the charge and momentum of hadrons in magnetic field. Indeed, this feature would allow to discriminate between tau neutrinos and anti-neutrinos also in the hadronic decay channels of the tau lepton. The emulsion target is complemented by high resolution tracking chambers to provide the time stamp to the event and connect muon tracks from the target to the muon spectrometer. The muon spectrometer is based on the concept developed for the OPERA apparatus: a dipolar iron magnet where high precision tracking chambers provide the momentum and coarse resolution chambers provide the tracking within the iron slabs. About 10,000 tau neutrino interactions are expected to be observed in SHiP.

The emulsion target also acts as the target of very weakly interacting particles, like the dark matter, produced at the accelerator, if their mass is in the GeV range. Unlike the non-relativistic galactic dark matter producing nuclear recoils of the keV energy range, dark matter produced at the accelerator is ultra-relativistic and it could be observed through its scattering off the electrons of the emulsion target of the neutrino detector. The elastic interaction of dark matter particles with electrons produces one electron in the final state, thus mimicking elastic interaction of neutrinos that constitute the main background for this search. In [171] the sensitivity

to light dark matter shows to be very competitive with all the planned experiments in the next decade.

The SHiP Collaboration is preparing a Comprehensive Design Report to be submitted within 2018, in the framework of the Physics Beyond Colliders working group, that will be evaluated within 2020. The construction and installation is expected to start in 2021 with data taking to start in 2026.

9.6.3 Projects in Applied Science

9.6.3.1 Muon Radiography

Muon radiography measures the absorption of cosmic-ray muons in matter, analogously to the conventional radiography that makes use of X-rays. The interaction of primary cosmic-rays with the atmosphere provides an abundant source of muons that can be used for various applications of muon radiography. Muon radiography was first proposed to determine the thickness of snow layers on a mountain [174]. The first application was realised in 1971 with the seminal work of Alvarez and collaborators searching for unknown burial cavities in Chephren's pyramid [175]. The pioneering work done in Japan for the radiography of the edifice of volcanoes by using quasi-horizontal cosmic-ray muons [176, 177] has opened new possibilities for the study of their internal structure.

Nuclear emulsions were used for the first time in 2006 for the muon radiography of the Asama volcano in Japan [178]. The main advantages of the emulsion technique are the simplicity and portability of the detector setup, and the absence of power supplies and electronic data acquisition systems, usually difficult to transport and operate on the summit of a volcano.

In 2012 an emulsion detector was installed on the Stromboli volcano to image its crater region. Despite of the strong influence that the crater area and the Sciara del Fuoco slope have on the volcanic dynamics of the Stromboli island, their internal structure is not well known because of the limited resolution of conventional geophysical methods. An emulsion detector of 0.73 m^2 surface was exposed there and took data for about 5 months in 2012. Emulsion films were exposed in the form of two doublets separated by 5 mm iron slabs intended to reject the background induced by the soft component of cosmic-rays. Figure 9.31 shows an excess in the rate of muons in the crater region that is interpreted in terms of a lower density region [179]. This excess lies in the region B of Fig. 9.31, the one where the detector is sensitive to density variations, while A denotes the free sky region. In the region C, instead, the average thickness is larger than 800 m, such that the rate is too low to appreciate density changes. The data analysis provided an image of the crater area of Stromboli with a resolution of about 10 m in the center of the target area. The observed muon excess larger than 30% indicates an average density decrease along the muon path down to 1.7 g/cm^3 with respect to the standard rock density

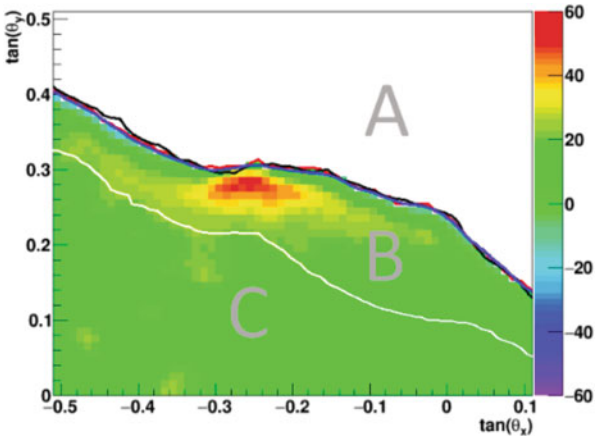


Fig. 9.31 Excess of the muon rate seen with an emulsion detector in the crater region of the Stromboli volcano [179]. The colour scale indicates the number of muons over an angular range of $10 \times 10\,\text{mrad}^2$ and a surface of $0.73\,\text{m}^2$

of $2.65\,\text{g/cm}^3$. Further measurements campaigns are foreseen with larger detector surfaces at Stromboli as well as on other volcanoes.

An interdisciplinary project between the fields of geosciences and particle physics was also initiated. This project aims to image the bases of Alpine glaciers in three dimensions via cosmic-ray muon radiography using emulsion particle detectors. The results will be used to test the models for erosional processes and provide clues revealing how the Alpine glaciers have been shaped. The results also have an impact on society since they can be used to check the possibility of disasters caused by glacier retreats. However, studying the morphology of active Alpine glaciers has been a difficult task due to the lack of technology. Muon radiography is considered as a powerful tool to address this issue. The technique has been applied to map the bases of the Eiger Glacier and Aletsch Glaciers in Central Swiss Alps, where the Jungfrau railway tunnel provides a situation suitable for placing the detectors.

Recently, a measurement at the upper part of the Aletsch Glacier has been performed [180]. Muon detectors made of emulsion films were installed at three sites along the wall of the Jungfrau railway tunnel running through the bedrock underneath the Aletsch Glacier. The detectors had a total effective detection area of $250\,\text{cm}^2$ for each site, and the data were collected for 47 days. The shape of the boundary between glacial ice and the bedrock at the upper part of the Aletsch Glacier was measured as shown in Fig. 9.32. This is the first successful application of this technology to a glaciated environment, which demonstrates that muon radiography can be a complementary method for determining the bedrock topography in such an environment when suitable detector sites are available. To image the bedrock topography underlying the Eiger glacier, another measurement is underway.

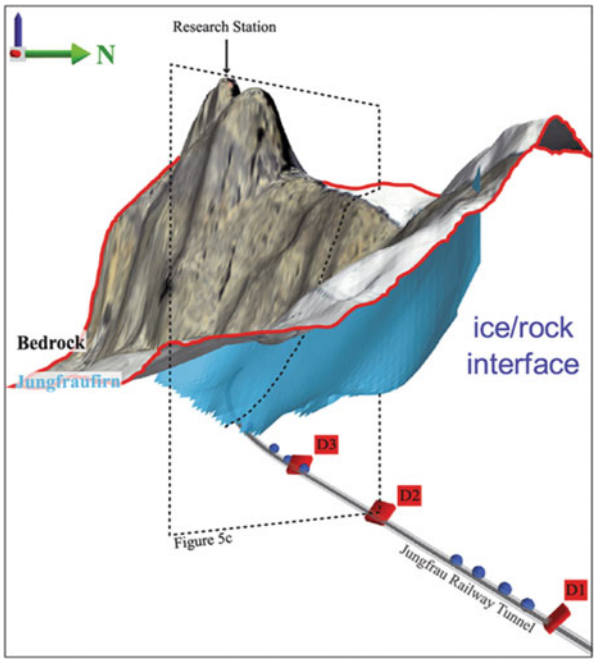


Fig. 9.32 The three-dimensionally reconstructed ice–rock interface (blue surface) determined via muon radiography analysis [180]

This technique has also been applied to other fields such as investigations of archaeological sites (pyramids and tumuli). One recent success was the discovery of a large void in Khufu's Pyramid [181]. This void was first observed with emulsion detectors installed in the Queen's chamber and later confirmed with scintillator hodoscopes placed in the same chamber and with gas detectors outside the pyramid. Figure 9.33 shows that large known structures were observed as expected. In addition, an unexpected muon excess was observed, indicating that there is an additional void. This discovery demonstrated that this technique is useful for such investigations. The muon radiography technique with emulsions has been established and is further broadening our knowledge in several new fields, such as safety inspections, by looking for underground cavities or diagnosing furnace problems.

9.6.3.2 Medical Applications

Medical applications of the emulsion technique have also been attempted in the last decade. In the treatment of cancer by hadron-therapy, beams of carbon nuclei present therapeutic advantages over proton beams. The knowledge of the fragmentation of carbon nuclei when they interact with human tissues is important to evaluate the spatial profile of the energy deposition in the human body, thus maximizing the effectiveness in hitting the cancer with minimal damage to the

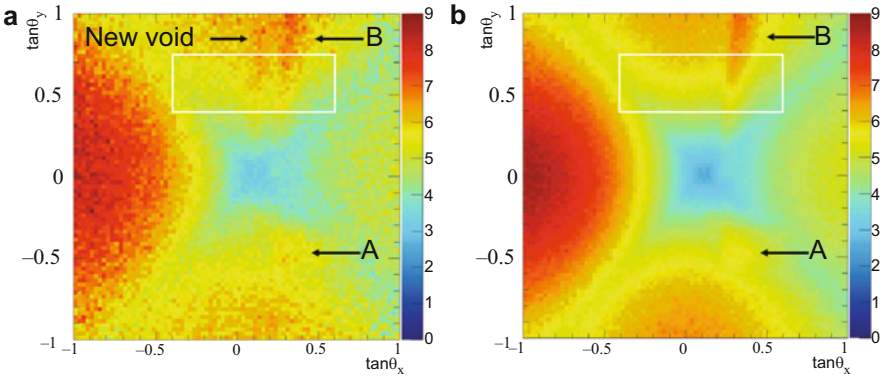


Fig. 9.33 Two-dimensional histogram of the detected muon flux at a position (left) and the result of a simulation with the known inner structures (right) [181]. The large known structures (A: the King's chamber and B: the Grand Gallery) and a new void were observed. The colour scale indicates muons per square centimetre per day per steradian

neighbouring tissues. For this purpose, ECC detectors simulating human tissues have been realized and exposed to ion beams. The ECC technique, in fact, allows to integrate target and tracking devices in a very compact structure. This fact, together with the development of techniques of controlled fading of particle tracks in nuclear emulsions [2], has opened the way to measurements of the specific ionization over a very broad dynamic range. The application of several refreshing treatments to the emulsion films makes them sensitive to different ionization values. The combined analysis of several films allows to overcome saturation effects, so that films normally sensitive to minimum ionizing particles can be used to measure the charge of carbon ions and of their induced fragments. Details of the technique are reported in [182, 183].

Figure 9.34 shows the identification of fragments produced by the interaction of carbon ions with Lexan plates, which simulate the human body tissues given the similar electron density [183]. Such a charge identification capability allowed the measurement of the charge-changing cross-section of carbon ions with water by placing the target ECC inside a water tank [184] as well as the charge-changing cross-section of carbon ions with lexan [185].

In the framework of the FIRST (Fragmentation of Ions Relevant for Space and Therapy) experiment [186], two Emulsion Cloud Chambers were exposed to the fragments produced by a ^{12}C beam (400 MeV/n) impinging on a composite target. The detectors were located in such a way to collect ^{12}C fragments emitted at large angles with respect to the beam axis, as shown in the left plot of Fig. 9.35. Indeed, the characterization of secondary fragments produced by a ^{12}C beam incident on a target is crucial to monitor the dose deposition inside the patient and to estimate the overall biological effectiveness due to the fragmentation of the incident beam. Data available in literature are rather scarce in this respect. Films used in this experiment were belonging to the same batch produced for the OPERA experiment.

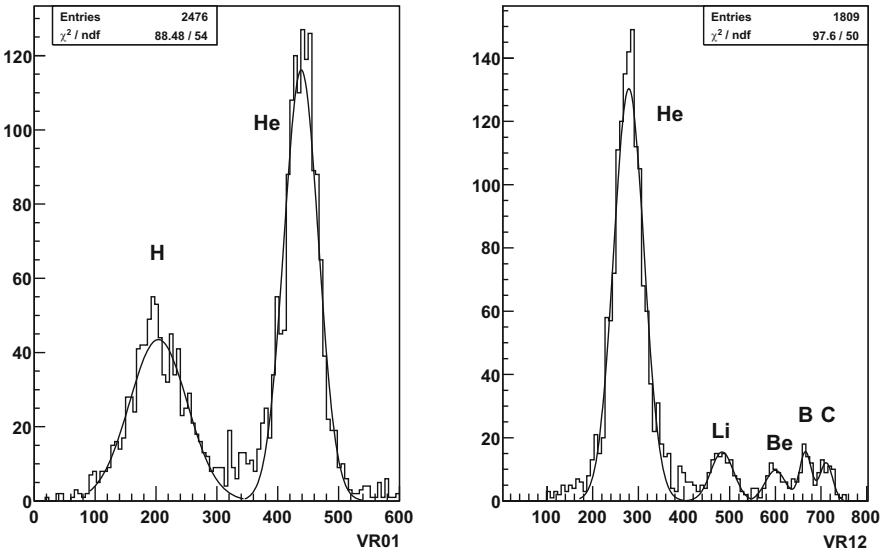


Fig. 9.34 Measurement of the electric charge of nuclear fragments produced by carbon ion interactions in an ECC detector [183]. Left: separation between hydrogen and helium ions. Right: separation of heavier fragments

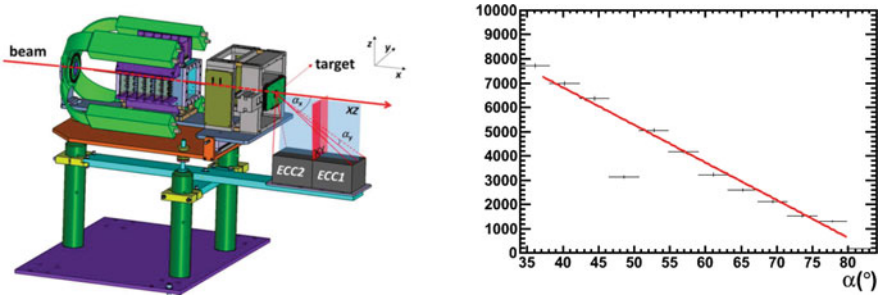


Fig. 9.35 Left: ECC used in the FIRST experiment. Right: angular distribution of protons [188]

The ECC structure was made of two sections [187]: the first section, consisting of six nuclear emulsion films, was meant to trigger all the incoming fragments entering the detector; the second section, consisting of 55 nuclear emulsion films interleaved with 1 mm thick lead plates, was optimised for the momentum measurements of fragments using the particle range. Given the peculiar geometry, tracks impinge on the emulsion films with rather large incident angles. Recent developments in the scanning technology [123, 141, 142] were essential to analyse these films. Almost 37,000 proton tracks were fully reconstructed and their angular and momentum spectra were measured in a wide angular range extending for the first time to more

than 80°, as shown in the right plot of Fig. 9.35. The momentum was also measured through the particle range as reported in [188].

The FOOT (FragmentatiOn Of Target) experiment [189] is designed to study target and projectile fragmentation processes. Target nuclei (^{16}O,^{12}C) fragmentation induced by 150–250 MeV proton beams will be studied via the inverse kinematic approach. The detector includes a magnetic spectrometer based on silicon pixel detectors and drift chambers, a scintillating crystal calorimeter with TOF capabilities, thick enough to stop the heavier fragments produced, and a ΔE detector based on scintillating bars to achieve the needed energy resolution and particle identification. An alternative setup of the experiment will exploit the emulsion chamber capabilities. Dedicated emulsion chambers will be coupled with the interaction region to measure the interaction vertices within the target, tag the produced light charged fragments such as protons, deuterium, tritium and Helium nuclei, and measure their angular and momentum spectra. Given the very good identification capability of the emulsion technology for low Z fragments, the results from the emulsion chamber detectors are expected to be of particular interest also for the radio-protection in the space where Helium is relevant. The FOOT data taking is foreseen at the CNAO centre in Pavia starting from 2018 with the emulsion setup, while the electronic detectors will start data taking in 2019. Data taking in the major European laboratories such as HIT at Heidelberg and GSI at Darmstadt is also foreseen.

References

1. Tadaaki Tani, Photographic Science, Advances in Nanoparticles, J-Aggregates, Dye Sensitization, and Organic Devices, Oxford University Press (2011), ISBN: 9780199572953.
2. T. Nakamura et al., Nucl. Instrum. Meth. **A556** (2006) 80.
3. C. Amsler et al., JINST **8** (2013) P02015.
4. P.H. Fowler, D.H. Perkins and C.F. Powell, The study of elementary particles by the photographic method, Pergamon Press (1959).
5. W.H. Barkas, Nuclear research emulsion, Academic Press, New York, 1973.
6. H. Becquerel, C.R. Academy of Science, **122** (1896) 501 and **122** (1896) 1086.
7. S. Kinoshita, Proc. Royal Society (London) **83A** (1910) 432.
8. C.F. Powell, G.P.S. Occhialini, D.L. Livesey and L.V. Chilton, Journal of Sci. Instrum. **23** (1946) 102.
9. M. Blau and J. A. De Felice, Phys. Rev. **74** (1948) 1198.
10. H. Yagoda, Phys. Rev. **79** (1950) 207.
11. P. Demers, Sci. and Indust. Phot. **23** (1952) 1.
12. B. Stiller, M. M. Shapiro and P.W. O'Dell, Rev. Sci. Instr. **25** (1954) 340.
13. C.F. Powell, Phil. Mag. **44** (1953) 219.
14. C.M.G. Lattes, H. Muirhead, G.P.S. Occhialini and C.F. Powell, Nature **159** (1947) 694.
15. C.M.G. Lattes, G.P.S. Occhialini and C.F. Powell, Nature **160** (1947) 453.
16. C.M.G. Lattes, G.P.S. Occhialini and C.F. Powell, Nature **160** (1947) 486.
17. W.F. Fry, J. Schneps, and M.S. Swami, Phys. Rev. **103** (1956) 1904–1905;
    S. Lokanathan, D.K. Robinson and S.J. St Lorant, Proc. of the Royal Society of London, Series A, Mathematical and Physical Sciences, **254**, No. 1279 (1960) 470.

18. N.A. Dobrotin, Sov. Phys. Usp. **2** (1960) 974;
M.F. Kaplon and D. M. Ritson, Phys. Rev. **88** (1952) 386.
19. M.F. Kaplon, B. Peters, H.L. Reynolds and D.M. Ritson, Phys. Rev. **88** (1952) 295.
20. J. Nishimura, Soryushiron Kenkyu **12** (1956) 24.
21. K. Niu, Proc. Jpn. Acad., Ser. **B 84** (2008) 1;
K. Niu, proc. of the I International Workshop on Nuclear Emulsion Techniques, Nagoya, June 1998.
22. C.M.G. Lattes et al., Prog. Theor. Phys. Suppl. **47** (1971) 1.
23. I. Ohta, K. Kasahara, I. Mito, A. Ohsawa, T. Taira and S. Torii, Proc. of the Int. Cosmic Ray Conf., Denver, **3** (1973) 2250;
M. Akashi et al., Proc. of the Int. Cosmic Ray Conf., Munich, **7** (1975) 2549.
24. A.V. Apanasenko et al., Astropart. Phys. **16** (2001) 13.
25. T.H. Burnett et al., Nucl. Instrum. Meth. **A251** (1986) 583.
26. K. Niu, E. Mikumo, Y. Maeda, Prog. Theor. Phys. **46** (1971) 1644.
27. J.E. Augustin et al., Phys. Rev. Lett. **33** (1974) 1406.
28. J.J. Aubert et al., Phys. Rev. Lett. **33** (1974) 1404.
29. S. Ogawa et.al, Soryushiron Kenkyu **43** (1971) 801;
Z. Maki and T. Maskawa, Prog. Theor. Phys. **46** (1971) 1647;
T. Hayashi et al., Prog. Theor. Phys. **47** (1972) 280.
30. K. Hoshino et al., Proc. Int. Cosmic Ray Symp. on High Energy Phenomena (1974) 161;
H. Sugimoto et al., Prog. Theor. Phys. **53** (1975) 1541.
31. K. Hoshino et al., Proc. of the Int. Cosmic Ray Conf., Munich, **7** (1975) 2442.
32. N. Ushida et al., Nucl. Instrum. Meth. **224** (1984) 50.
33. E.H.S. Burhop et al., Phys. Lett. **B65** (1976) 299.
34. D. Allasia et al., Nucl. Phys. **B176** (1980) 13.
35. C. Angelini et al., Phys. Lett. **B84** (1979) 150.
36. D. Allasia et al., Phys. Lett. **B87** (1979) 287.
37. K. Niwa, Physics and Astrophysics of Neutrinos, Springer, Berlin, M. Fukugita and A. Suzuki (Eds.), (1994) 520.
38. N. Ushida et al., Phys. Lett. **B206** (1988) 375.
39. T. Bolton, hep-ex/9708014.
40. G. De Lellis, P. Migliozzi and P. Santorelli, Physics Reports **399** (2004) 227.
41. N. Ushida et al., Phys. Lett. **B206** (1988) 380.
42. N. Ushida et al., Phys. Rev. Lett. **45** (1980) 1049;
N. Ushida et al., Phys. Rev. Lett. **45** (1980) 1053;
N. Ushida et al., Phys. Rev. Lett. **48** (1982) 844;
N.Ushida et al., Phys. Rev. Lett. **56** (1986) 1767.
43. N. Ushida et al., Phys. Rev. Lett. **47** (1981) 1694;
N. Ushida et al., Phys. Rev. Lett. **57** (1986) 2897.
44. S.W. Herb et al., Phys. Rev. Lett. **39** (1977) 252.
45. S. Aoki et al., Nucl. Instrum. Meth. **A274** (1989) 64.
46. K. Hoshino and G. Rosa, Nucl. Tracks Radiat. Meas. **12** (1986) 477.
47. S. Aoki et al., Nucl. Tracks and Rad. Meas. **12** (1986) 249.
48. J.P. Albanese et al., Phys. Lett. **B158** 186.
49. S. Aoki et al., Prog. Theor. Phys., **89** (1993) 131.
50. K. Kodama et al., Nucl. Instrum. Meth. **A289** (1990) 146.
51. S. Aoki et al., Nucl. Instrum. Meth. B **51** (1990) 466.
52. K. Kodama et al., Nucl. Instrum. Meth. **B93** (1994) 340.
53. K. Kodama et al., Prog. Theor. Phys. **89** (1993) 679.
54. A. Shor, Phys. Lett. **B215** (1988) 375.
55. S. Aoki et al., Phys. Lett. **B224** (1989) 441.
56. N. Armenise et al., Nucl. Instrum. Meth. **A361** (1995) 497.
57. R. Tanaka and N. Ushida, Nucl. Phys. **A585** (1995) 323;
S. Aoki et al., Phys. Rev. Lett. **65** (1990) 1729.

58. M. Danysz et al., Nucl. Phys. **49** (1963) 121;
D. Prowse, Phys. Rev. Lett. **17** (1966) 782.
59. S.Aoki et al., Prog. Theor. Phys. **85** (1991) 951;
S.Aoki et al., Prog. Theor. Phys. **85** (1991) 1287.
60. H. Takahashi et al., Phys. Rev. Lett. **87** (2001) 212502.
61. K. Nakazawa, Nucl. Phys. **A585** (1995) 75.
62. H. Takahashi et al., Nucl. Phys. **A721** (2003) 951.
63. A. Ichikawa et al., Phys. Lett. **B500** (2001) 37.
64. T. Nakazawa, J. Soc. Photogr. Sci. Technol. Japan **71** No. 4 (2008) 245;
E. Hayata et al., P07 Proposal to J-PARC (2006).
65. K. Niwa, K. Hoshino and K. Niu (1974) Proc. Int. Cos. Ray Simp. on High Energy Phenomena. (Cos. Ray Lab., Univ. of Tokyo) (1974) 149.
66. T. Nakano, PhD Thesis, University of Nagoya (1997).
67. G. Rosa et al., Nucl. Instrum. Meth. A **394** (1997) 357.
68. K. Kodama et al., Nucl. Instrum. Meth. **A493** (2002) 45.
69. T. Nakano, Butsuri **56** (2001) 411.
70. M. De Serio et al., Nucl. Instrum. Meth. **A554** (2005) 247.
71. M. De Serio et al., Nucl. Instrum. Meth. **A512** (2003) 539.
72. K. Kodama et al., Nucl. Instrum. Meth. **A574** (2007) 192.
73. K. Kodama et al., Rev. Sci. Instrum. **74** (2003) 53.
74. L. Arrabito et al., JINST **2** (2007) P02001.
75. E. Eskut et al., Nucl. Instrum. Meth. A **401** (1997) 7.
76. S. Aoki et al., Nucl. Instrum. Meth. A **447** (2000) 361.
77. E. Eskut et al., Nucl. Phys. **B793** (2008) 326.
78. A. Kayis-Topaksu et al., Phys. Lett. **B539** (2002) 188.
79. G. Onengut et al, Phys. Lett. **B613** (2004) 105;
A. Kayis-Topaksu et al., Phys. Lett. **B555** (2003) 156;
A. Kayis-Topaksu et al., Phys. Lett. **B575** (2003) 198.
80. A. Kayis-Topaksu et al., Eur. Phys. J. **C52** (2007) 543.
81. G. De Lellis et al., Nucl. Phys. **B763** (2007) 268.
82. K. Kodama, et al., FERMILAB-PROPOSAL-0803 (Oct 1993).
83. A. Ereditato, G. Romano and P. Strolin, Nucl. Phys. **54B** (1997) 139.
84. A.S.Ayan et al., CERN-SPSC/97-5, SPSC/I213 (1997).
85. K. Kodama et al., Nucl. Instrum. Meth. **A516** (2004) 21.
86. K. Kodama et al., Phys. Lett. **B504** (2001) 218.
87. K. Kodama et al., Phys. Rev. **D78** (2008) 052002.
88. T. Nakano, J. Soc. Photogr. Sci. Technol. Japan **71** No. 4 (2008) 229.
89. N. Armenise et al., Nucl. Instrum. Meth. **A551** (2005) 261.
90. L. Arrabito et al., Nucl. Instrum. Meth. **A568** (2006) 578.
91. L. Arrabito et al., JINST **2** (2007) P05004.
92. I. Kreslo et al., JINST **3** (2008) P04006.
93. K. Borer et al., Nucl. Instrum. Meth. **A566** (2006) 327.
94. Y. Fukuda et al., Phys. Rev. Lett. **81**, 1562 (1998);
W.W.M. Allison et al., Phys. Lett. **B449** (1999) 137;
M. Ambrosio et al., Phys. Lett. **B517** (2001) 59.
95. M.H. Ahn et al., Phys. Rev. **D74**, 072003 (2006);
D.G. Michael et al., Phys. Rev. Lett. **97** (2006) 191801.
96. K. Niwa, Contrib. to the "Snowmass '94" conference on particle and nuclear astrophysics and cosmology in the next millennium, Snowmass, 1994.
97. A. Ereditato, K. Niwa and P. Strolin, INFN/AE-97/06, Nagoya DPNU-97-07, Jan 27th 1997.
98. A. Ereditato, K. Niwa and P. Strolin, Nucl. Phys. Proc. Suppl. **66** (1998) 423.
99. M. Guler et al., CERN-SPSC-2000-028.
100. R. Acquafredda et al., JINST **4** (2009) P04018.
101. A. Anokhina et al., JINST **3** (2008) P07005.

102. S. Miyamoto et al., Nucl. Instrum. Meth. **A575** (2007) 466.
103. N. Agafonova et al. [OPERA Collaboration], Eur. Phys. J. **C74** (2014) no.8, 2986.
104. A. Kayis-Topaksu et al., New J. Phys. **13** (2011) 093002.
105. N. Agafonova et al. [OPERA Collaboration], New J. Phys. **14** (2012) 013026.
106. N. Agafonova et al. [OPERA Collaboration], Phys. Lett. B691 (2010) 138.
107. N. Agafonova et al. [OPERA Collaboration], JHEP 1311 (2013) 036.
108. N. Agafonova et al. [OPERA Collaboration], Phys. Rev. **D89** (2014) no.5, 051102.
109. N. Agafonova et al. [OPERA Collaboration], PTEP 2014 (2014) no.10, 101C01.
110. N. Agafonova et al. [OPERA Collaboration], Phys. Rev. Lett. 115 (2015) no.12, 121802.
111. H. Ishida et al., PTEP 2014 (2014) N. 9, 093C01.
112. A. Longhin et al., IEEE Trans. Nucl. Sci. **62** (2015) 2216–2225.
113. C. Patrignani et al., Review of Particle Physics, Chin. Phys. **C40** (2016) N. 10, 100001.
114. N. Agafonova et al., [OPERA Collaboration], Physical Review Letters 120 (2018) 211801.
115. N. Agafonova et al. [OPERA Collaboration], JHEP **07** (2013) 004.
116. N. Agafonova et al., [OPERA Collaboration], arXiv:1803.11400
117. N. Agafonova et al. [OPERA Collaboration], JHEP 1506 (2015) 069.
118. NVIDIA CUDA web page: http://www.nvidia.com/cuda.
119. T. Fukuda et al., JINST 9 (2014) P12017.
120. A. Ariga and T. Ariga, JINST **9** (2014) P04002.
121. A. Alexandrov, et al., JINST **10** (2015) P11006
122. M. Yoshimoto, T. Nakano, R. Komatani and H. Kawahara, PTEP **10** (2017) 103.
123. A. Alexandrov et al., Nature Scientific Reports **7** (2017) 7310.
124. N. Natsume al., Nucl. Instr. Meth. A **575** (2007) 439.
125. T. Naka et al., Nucl. Instrum. Meth. A **718** (2013) 519.
126. T. Ariga et al., JINST **11** (2016) P03003.
127. A. Nishio et al., PoS KMI2017 (2017) 057.
128. J. Nishimura, Adv. Space Res. **30** No. 5 (2002) 1071..
129. K. Kodama et al., Adv. Space Res. **37** (2006) 2120.
130. S. Aoki et al., J. Soc. Photogr. Sci. Technol. Japan **71** No. 4 (2008) 256.
131. K. Ozaki et al., Nucl. Instrum. Meth. **A833** (2016) 165.
132. S. Takahashi, S. Aoki for GRAINE collaboration, Adv. Space Res. (in press), https://doi.org/10.1016/j.asr.2017.08.029.
133. S. Takahashi et al., PTEP 2015 **4** (2015) 043H01.
134. The P7SOURCE_V6 Instrument Response Functions, https://www.slac.stanford.edu/exp/glast/groups/canda/archive/pass7v6/lat_Performance.htm.
135. S. Takahashi et al., Nucl. Instrum. Meth. **A620** (2010) 192.
136. H. Rokujo et al., Nucl. Instrum. Meth. **A701** (2013) 127.
137. S. Takahashi et al., PTEP 2016 **7** (2016) 073F01.
138. K. Ozaki et al., JINST **10** (2015), P12018.
139. J. B. R. Battat et al., Physics Reports **662** (2016) 1.
140. A. Alexandrov et al., LNGS-LOI 48/15, arXiv:1604.04199.
141. A. Alexandrov et al., JINST **10** (2015) P11006.
142. A. Alexandrov et al., JINST **11** (2016) P06002.
143. T. Naka et al., EAS Publ. Ser. **53** (2012) 51.
144. M. Kimura and T. Naka, Nucl. Instrum. Meth. A **680** (2012) 12.
145. H. Tamaru et al., Applied Phys. Lett. **80** (2002) 1826.
146. A. Alexandrov et al., Astroparticle Physics **80** (2016) 16.
147. N. Naganawa et al., Phys. Procedia **88** (2017) 224.
148. K. Mishima et al., Nucl. Instrum. Meth. **A600** (2009) 342.
149. N. Naganawa et al., PoS KMI2017 (2017) 077.
150. N. Naganawa, T. Ariga, S. Awano, M. Hino, K. Hirota, H. Kawahara, M. Kitaguchi, K. Mishima, H. M. Shimizu, S. Tada, S. Tasaki, A. Umemoto, arXiv:1803.00452.
151. T. Nakazawa, Few-Body Systems, 2013, Volume 54, Issue 7-10, pp1279–1282.
152. G. Drobychev et al., AEgIS Proposal, CERN-SPSC-P-334 (2007).

153. S. Aghion et al. [AEgIS Collaboration], JINST 8 (2013) P08013.
154. The Antiproton Decelerator webpage, https://home.cern/about/accelerators/antiproton-decelerator.
155. T. Ariga et al. [AEgIS Collaboration], Int. J. Mod. Phys. Conf. Ser. 30 (2014) 1460268.
156. S. Aghion et al. [AEgIS Collaboration], Nature Communications 5 (2014) 4538.
157. S. Aghion et al. [AEgIS Collaboration], JINST 12 (2017) P04021.
158. A. Ferrari, P.R. Sala, A. Fasso and J. Ranft, CERN-2005-010, CERN, Geneva Switzerland, (2005).
159. T.T. Bohlen et al., Nucl. Data Sheets 120 (2014) 211.
160. S. Aghion et al., JINST 11 (2016) P06017.
161. K. Suzuki et al., Prog. Theor. Exp. Phys. (2015) 053C01.
162. K. Abe et al. [Hyper-Kamiokande Proto-Collaboration], PTEP 2015 (2015) 053C02.
163. R. Acciarri et al. [DUNE Collaboration], FERMILAB-DESIGN-2016-01 (2016).
164. K. Yamada et al., PTEP 2017 **6** (2017) 063H02.
165. K. Abe et al. [T2K Collaboration], Nucl. Instrum. Meth. **A 659** (2011) 106.
166. T. Fukuda et al., PTEP 2017 **6** (2017) 063C02.
167. T. Fukuda on behalf of the NINJA Collaboration, PoS KMI2017 (2017) 012.
168. S. Aoki et al. [DsTau Collaboration], CERN-SPSC-2017-029, SPSC-P-354 (2017).
169. A. Akmete et al. [SHiP Collaboration], CERN-SPSC-2017-033, SPSC-EOI-017 (2017).
170. H. Dijkstra and T. Ruf, CERN-SHiP-NOTE-2015-009. http://cds.cern.ch/record/2115534/files/SHiP-NOTE-2015-009.pdf
171. M. Anelli et al., CERN-SPSC-2015-016, SPSC-P-350, arXiv:1504.04956.
172. S. Alekhin et al., Rept. Prog. Phys. **79** (2016) no.12, 124201.
173. A. Akmete et al., JINST **12** (2017) no.05, P05011.
174. E.P. Georg, Commonw. Eng., July 1955.
175. L.W. Alvarez et al., Science **167** (1970) 832.
176. K. Nagamine et al., Nucl. Instrum. Meth. **A356** (1995) 585.
177. H.K.M. Tanaka et al., Nucl. Instrum. Meth. **A507** (2003) 657;
H.K.M. Tanaka et al., Nucl. Instrum. Meth. **A555** (2005) 164.
178. H.K.M. Tanaka et al., Earth and Planetary Science Letters **263** (2007) 104;
H.K.M. Tanaka et al., Nucl. Instrum. Meth. **A575** (2007) 489;
H.K.M. Tanaka et al., Geophysical Research Letters **34** (2007) 389.
179. V. Tioukov et al., Annals of Geophysics 60 (2017) N. 1 S0111.
180. R. Nishiyama, A. Ariga, T. Ariga, S. Kaser, A. Lechmann, D. Mair, P. Scampoli, M. Vladymyrov, A. Ereditato, F. Schlunegger, Geophysical Research Letters 55933 (2017).
181. K. Morishima et al., Nature **552** (2017) 386–390.
182. T. Toshito et al., Nucl. Instrum. Meth. **A556** (2006) 482.
183. G. De Lellis et al., JINST **2** (2007) P06004.
184. T. Toshito et al., Phys. Rev. **C75** (2007) 054606.
185. G. De Lellis et al., Nuclear Physics A **853** (2011) 124–134.
186. R. Pleskac et al., Nucl. Instrum. Meth. **A678** (2012) 130–138.
187. A. Alexandrov et al., Meas. Sci. Technol. **26** (2015) 094001.
188. A. Alexandrov et al., JINST **12** (2017) P08013.
189. S. Argiro et al., The 26th International Nuclear Physics Conference, 11–16 September, 2016, Australia, Proceedings of Science, INPC2016, 128.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 10
Signal Processing for Particle Detectors

V. Radeka

10.1 Introduction

This chapter covers the principles and basic limits of signal processing for detectors based on measurements of charge induced on predominantly capacitive electrodes. While this presents a very limited scope, it already includes a broad range of different detector technologies employed in experiments in several areas of science and in various imaging devices. Detector technologies of interest involve semiconductors, gas and liquid ionization media as well as photo detectors converting (scintillation light) into photo-electrons or ionization. One class of detectors not considered here are bolometric detectors (see Sect. 10.4).

The literature cited in this chapter is twofold: the textbooks, tutorials and review articles which may serve to provide a systematic introduction to the reader, and references to journal articles describing specific applications and technological solutions. The former, while very good and useful, are unfortunately few, the latter are only a small selection from the vast body of journal articles and conference records. The former are listed first [1–10], and the latter are cited along with the material presented in this chapter.

For many detectors, particularly very large scale detectors used at colliding beam machines in particle and nuclear physics, systems aspects require most of the attention in the design. In high precision measurements of energy, time arrival or position of the incident particle or photon the noise introduced in the measurement is of primary interest. Each area of science may impose greatly different requirements on various performance parameters of the detector and signal processing. A silicon pixel detector for particle tracking at a high luminosity collider requires very short

V. Radeka (✉)
Brookhaven National Laboratory, Upton, NY, USA
e-mail: radeka@bnl.gov

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_10

pulse shaping (a few tens of nanoseconds) and it can tolerate a noise level of several hundred electrons rms. Consequently, a leakage (dark) current contributing shot noise may be 1 nA/cm^2 or more. A silicon detector for x-ray spectroscopy in photon science must be read out with a total noise of less than ten electrons rms with a shaping time of the order of 1 μs. This allows a leakage current of only ~10 pA/cm^2 or less. This chapter should enable the reader to evaluate the relations among such detector and readout parameters.

Signal processing for particle detectors rests on understanding of signal formation and of the sources of noise and their effects on measurement accuracy. Signal formation in detectors is based on electrostatics and it is calculated relatively easily starting from the Shockley-Ramo theorem [11, 12]. It gets more involved in multi-electrode detectors [13] and in crosstalk analysis. Signal processing with time-invariant systems has been extensively covered in the literature and is well understood.

Most innovations in signal processing in recent years have been in circuit implementations using monolithic CMOS technology. This technology has brought about a significant shift in the circuit concepts to time variant circuits due to the use of switched capacitance circuits for which CMOS transistors are well suited.

The noise analysis of time variant circuits brings up the question of whether to perform the analysis in frequency domain or in time domain. Analysis in frequency domain provides sufficient insight into time invariant circuits and it has been used in most of the literature. Frequency domain is less well suited for time variant circuit analysis as it does not provide much insight into the system transfer function. This is where the concept of the weighting function and Campbell's theorem [14] provide the tools which are simpler to use and provide more insight. While both analytical methods provide the same results in noise calculations, we note that particle detector signals are best described and are observed in the time domain, and so is the system response and the weighting function. In contrast, the noise analysis of narrow band circuits is best done in the frequency domain. The time domain analysis is based on the representation of noise as a random sequence of elementary impulses. In spite of our thinking and observing in the time domain, the device and circuit noise sources are customarily characterized in the frequency domain, e.g., we talk about the "white noise", "1/f noise", etc. Thus we switch our thinking between the two (Fourier transform related) domains depending on which one provides better insight and is easier to analyze in a particular case.

This chapter is intended to provide some insight into detector signal processing. Detailed circuit design, particularly of the monolithic circuits, has been rapidly developing and there has been a proliferation of publications. Advanced simulation tools are being used for noise analysis. Sometimes such an analysis provides numerical results without providing much insight into the role of various noise sources and into the overall weighting function of the signal processing chain. We concentrate here on the interpretation of the weighting function and on those aspects of signal processing and noise that have been less covered in the literature such as the induced signals in multi-electrode (strip and pixel) detectors, the "*kTC*" noise, correlated sampling and basic properties of low noise charge amplifiers. Induced

signals are determined using the "weighting field" concept, and noise analysis is based on the "weighting function" concept. The former is based on electrostatics and the latter on superposition of noise impulse contributions to the variable (current, charge) measured by the readout system.

10.2 Charge Collection and Signal Formation in Detectors

10.2.1 Current Induced by the Moving Charge and the Weighting Field Concept

Figure 10.1 illustrates the Shockley-Ramo theorem for induced signals, current and charge. $\boldsymbol{E_w}$ is the weighting field in units of 1/cm, and it is a measure of electrostatic coupling between the moving charge and the sensing electrode. The procedure to calculate the induced current as a function of time is as follows. First, the weighting field is determined by solving Poisson's equation analytically or numerically assuming unity potential on the sensing electrode of interest and zero potential on all other electrodes. Next, the velocity of the moving charge, $v = \mathrm{d}x/\mathrm{d}t$,

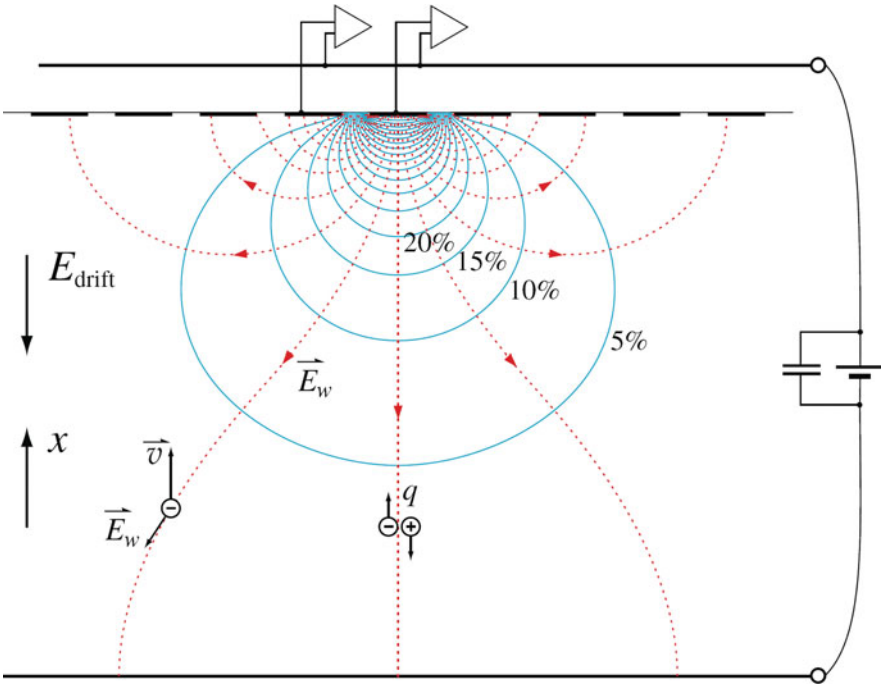


Fig. 10.1 Weighting potential (blue lines) and weighting field lines for planar strip electrode readout

as a function of position is determined from the operating (applied) field on the detector. This gives the induced current as a function of the position of the moving charge,

$$i = -q\,\vec{E}\cdot\vec{v}\,. \tag{10.1}$$

Third, the position of the moving *charge* as a function of time is determined by solving the equations of motion. This is necessary in the case of ballistic motion of charge, but it is simple in the case of transport by drift as the charge carriers follow the applied electric field. If we are interested only in the total induced charge and not in the waveforms, the induced charge is simply given by the difference in the weighting potentials between any two positions of the moving charge,

$$\begin{aligned} Q_s &= \int i\,dt = +\int \vec{E}_w\,d\vec{x}\,, \\ Q_s\,(1,2) &= q\,(V_{w2} - V_{w1})\,. \end{aligned} \tag{10.2}$$

An example of the weighting-field (potential) profiles is illustrated by the plot of equipotential lines for planar geometry with a strip sensing electrode. The operating (applied field) in this case is uniform and perpendicular to the electrodes. The weighting field map is in general quite different from that of the operating field; the two field maps are identical only in some special cases. The minus sign in Ramo's equation (Eq. 10.1) for the induced current results from the arbitrary assumption of induced current *into* the electrode being positive.

The sketch in Fig. 10.2a shows conceptually how the weighting field (potential) is defined: the sensing electrode is connected to unity potential, and all other electrodes to zero potential. The equipotential lines in Fig. 10.2b illustrate the solution for this case, showing two strips next to the sensing electrode. A great variety of results for the induced current and charge may arise in an electrode structure, such as this, depending on the particle type detected (distribution of ionization) and on the ratio of the charge observation measurement (or integration) time and the charge carrier transit time. The current waveforms shown are drawn qualitatively for a simple example. The operating field is assumed uniform and perpendicular to the electrode planes. Charge q_m traversing the full distance between the electrodes along line 1 is observed as $Q_1 = -q_m$, while the current decreases with distance from the sensing electrode 1, as the electrostatic coupling decreases. For a charge moving along line 2, the induced charge (i.e., the difference between the weighting potentials) is zero, if the measurement time is longer than the transit time. For a short measurement time a net induced charge is observed. The induced current waveform (the "crosstalk signal") is bipolar, since the weighting field direction changes along the path.

Figure 10.3 illustrates a simple case where the real (operating) electric field and the weighting field have the same form $= 1/d$. The induced current waveforms shown are for a semiconductor detector with different electron and hole mobilities. For extended ionization the waveforms result from superposition of the waveforms for localized ionization, and the currents decrease as the carriers arrive at the electrodes from different initial positions within the bulk of the detector.

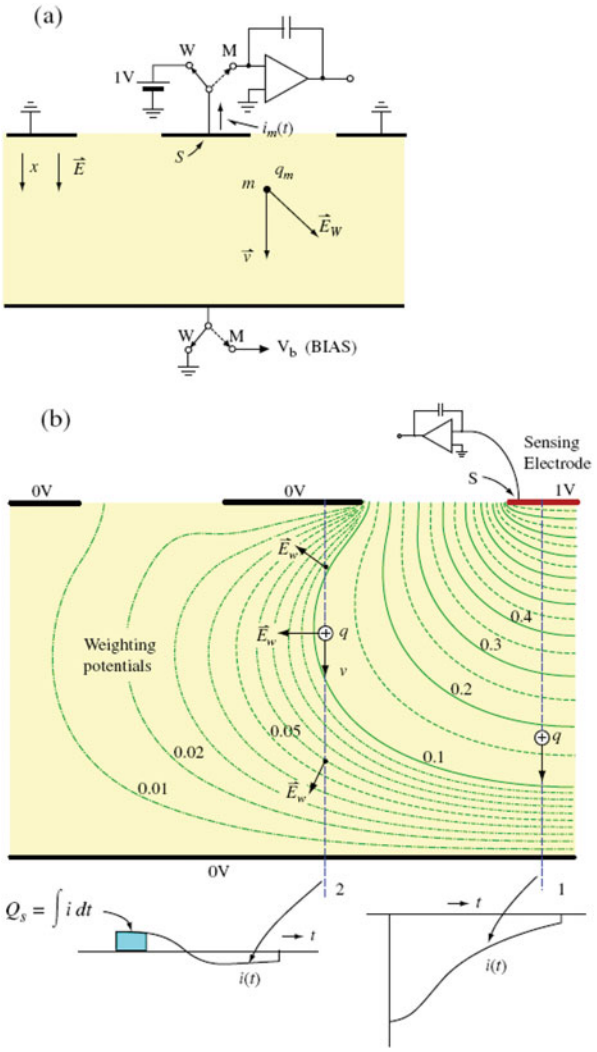


Fig. 10.2 Definition of the weighting potential: Solution of the Laplace equation for unity potential at the sensing electrode and zero potential at all other electrodes. From Radeka [9] Annual Reviews, www.annualreviews.org, by permission

10.2.2 Induced Current and Charge in Strip and Pixel Electrodes: Shielding Effect

The shielding effect is proportional to the ratio of the distance between the planar electrodes and the strip width (i.e., pixel radius). The shielding effect is more pronounced for pixels than for strips. The result of these configurations is that the

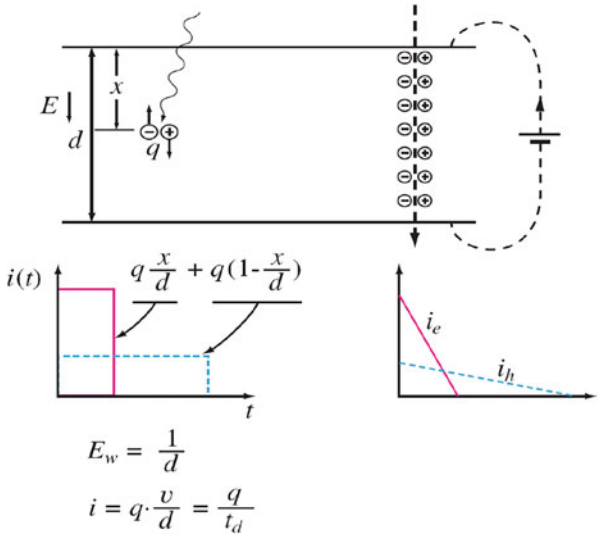


Fig. 10.3 Induced currents in infinite planar electrodes for localized and extended ionization

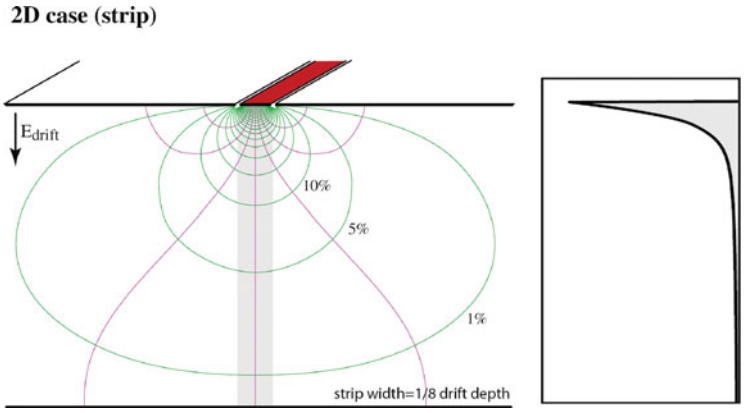


Fig. 10.4 Weighting field (potential) for strip electrode configuration

signal charge (integral of the induced current) is independent of the position of the origin of ionization for most of the volume of the detector except near the readout electrodes. This effect is used in detectors where only electrons are collected during the integration time, such as Cadmium Zinc Telluride (CZT), and some gas and noble liquid detectors. To illustrate this, histograms are shown in Figs. 10.4 and 10.5 for a strip and pixel illuminated by a beam of penetrating x-rays absorbed uniformly through the detector.

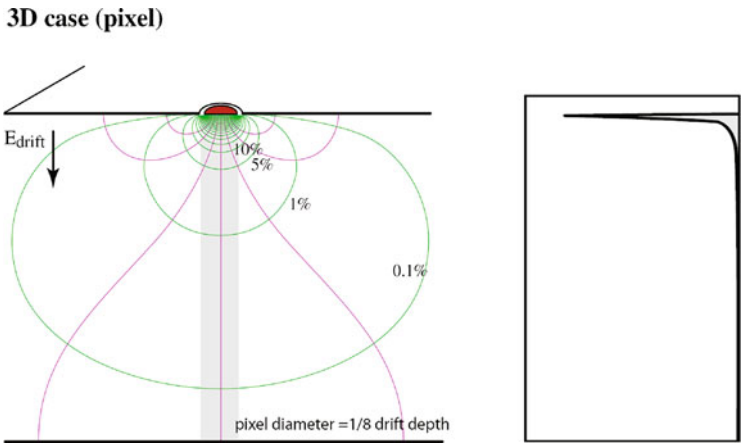


Fig. 10.5 Weighting field (potential) for pixel electrode configuration

10.2.3 Weighting Potential and Induced Charge in Co-Planar Electrodes

Coplanar grid readout was introduced for unipolar charge sensing by Luke [15] and it is commonly used with Cadmium Zinc Telluride (CZT) detectors. In such materials only electrons are collected (from the ionization produced by gamma rays or x-rays), the holes suffering from very low mobility and trapping. With parallel plane electrodes the induced charge for single carrier collection is dependent on the position (depth) where the ionization took place. In the coplanar grid concept one set of alternate strips is biased slightly more positively with respect to the other set of strips. This results in a drift field such that the signal electrons are collected on one set of strips only, Fig. 10.6. The weighting potential (field) for both sets of strips is identical, Fig. 10.7. The induced charges and currents are quite different, Figs. 10.8 and 10.9. Their respective differences are independent of the position, as can be concluded by following the weighting potential plot from any point on the planar sloped part of the plot to unity weighting potential for the collecting electrode, and (across the saddle) to zero potential for the non-collecting electrode, Fig. 10.9.

10.3 Noise: Origin and Properties

10.3.1 Noise Process and Noise Variance

The basis of a noise process can be represented as a sequence of randomly generated elementary impulses that has a Poisson distribution in time and mean rate of occurrence $\langle n \rangle$. Upon acting on a physical system with impulse response much

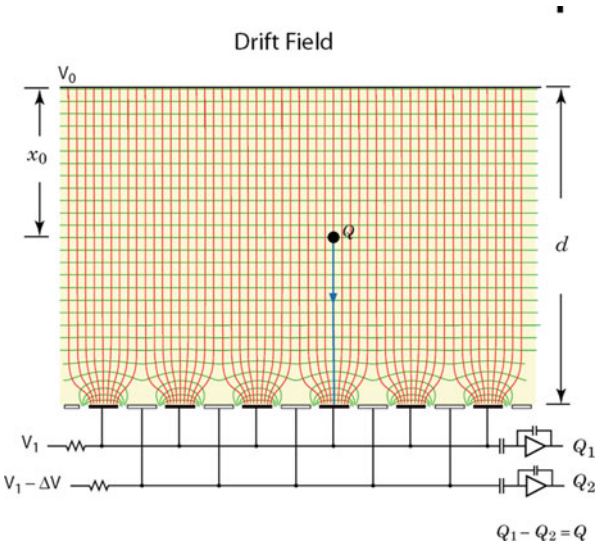


Fig. 10.6 Drift field for coplanar electrodes

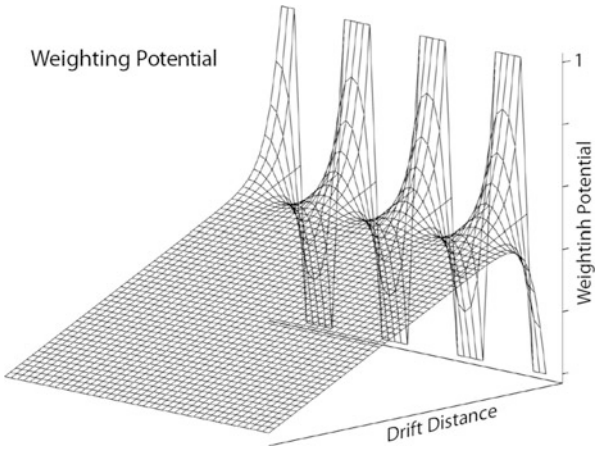


Fig. 10.7 Weighting potential for coplanar electrodes

longer than $\langle n \rangle^{-1}$, the characteristic noise waveforms (e.g., such as those we observe on an oscilloscope) are produced as a superposition of responses to individual impulses. The noise variance at the output of the physical system (a simple RC filter or a complete readout system) is calculated by using Campbell's theorem [14], which states that the sum of mean square contributions of all preceding impulses equals the variance. The expressions for the variance are given after subtracting the mean value. The variance is determined by the rate of impulses $\langle n \rangle$, their area q

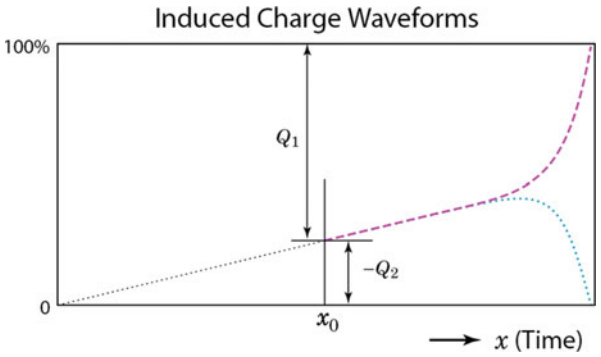


Fig. 10.8 Induced charges in coplanar electrodes: $Q_1 - Q_2 = const$ independently of x_0

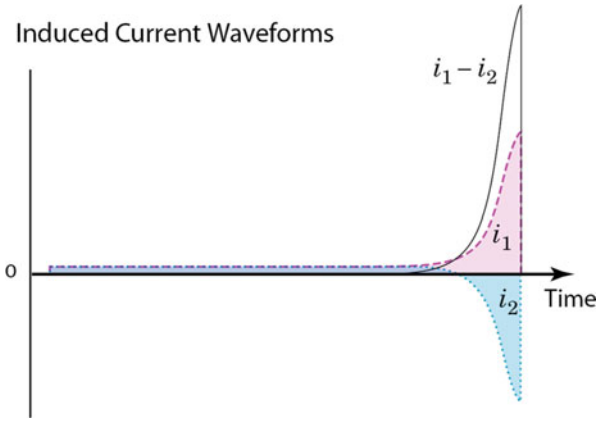


Fig. 10.9 Induced currents in coplanar electrodes

(charge), and by the impulse response $h(t)$, i.e., the weighting function $w(t)$ of the measurement system, the preamplifier and the subsequent readout chain,

$$\sigma^2 = \langle n \rangle \, q^2 \int_{-\infty}^{\infty} h^2(t)dt = \langle n \rangle \, q^2 \int_{-\infty}^{\infty} w^2(t)dt \; . \tag{10.3}$$

The noise variance is determined by the noise process, the rate of impulses $\langle n \rangle$, and their area q (charge), and by the impulse response $h(t)$. If we measure the variance and the $h(t)$, we can determine $\langle n \rangle q^2$, but we cannot determine $\langle n \rangle$ and q. Only when randomly generated carriers move in one direction, which results in a mean current $I_0 = \langle n \rangle q$, can the rate and the charge of impulses be determined from $\sigma^2$and I_0. It is shown in Ref. [9] that $\langle n \rangle q^2$ equals the mathematical (two-sided) noise current spectral density, whereas $\overline{i_n^2} = 2 \langle n \rangle \, q^2$equals the physical (single-sided) one, to be used in calculations of the equivalent noise charge (ENC) in Sect. 10.4.2.

10.3.2 A Model for Generation of Noise Spectra

Almost any noise spectrum can be generated from a random sequence of impulses (i.e., white noise with "infinite bandwidth") by using an appropriate filter, as illustrated in Fig. 10.10. These impulses may be either of only one polarity or of both polarities (current thermally generated in a p-n junction under reverse bias in the former, and with zero bias in the latter case). The mean value depends on the impulse polarities, but the variance does not.

"Infinite bandwidth" implies a noise spectrum which is flat over the frequency range where our measurement system has a non-zero response. Simple integration of white noise results in "random walk" with $1/f^{2}$ spectrum. An elementary impulse response for generation of this noise is the step function $U(t)$. Generation of $1/|f|$ noise is somewhat more elaborate. It requires fractional integration of *order one half*. The impulse response of the transforming filter is $U(t)/t^{1/2}$, as shown in the figure. The basic feature of any noise generating mechanism for low frequency divergent noises is an "infinitely long memory", i.e., very long memory, for individual independent elementary perturbations.

For a discussion of the basics of power-law spectra and of fractional integration see Ref. [16].

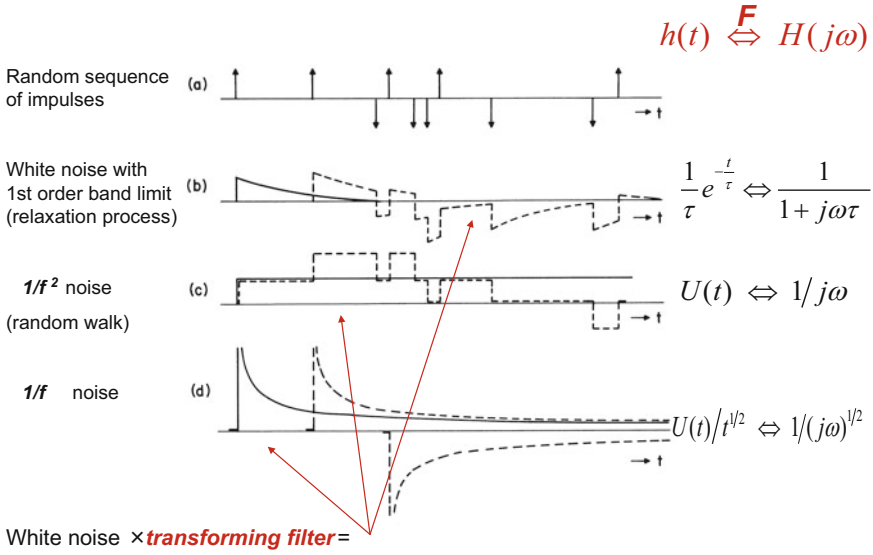


Fig. 10.10 Generation of some basic noise spectra from white noise by a transforming filter

10.3.3 Random Telegraph Noise and 1/f Noise

A noise spectrum very close to $1/|f|$ can be generated by superposition of relaxation processes with uniform distribution of life times, as illustrated in Fig. 10.11 The relaxation process is described by $U(t)\exp(-t/\tau)$, which represents a step change with exponential decay. Trapping-detrapping in semiconductors is one such possible mechanism for generation of $1/|f|$ noise. Since a simple RC integrator has the same response, a hardware filter which transforms white noise into $1/|f|$ noise can be made requiring about one time constant (one RC circuit) per decade of frequency, as shown in Ref. [16].

A single trap in a very small (minimum size) MOS transistor results in a drain current modulation known as random telegraph noise (RTS). This noise presents a limit to sensitivity in imaging arrays with a pixel capacitance of a few femtofarads and other noise sources reduced to a few electrons rms. There is extensive literature on RTS, e.g., Refs. [17–20].

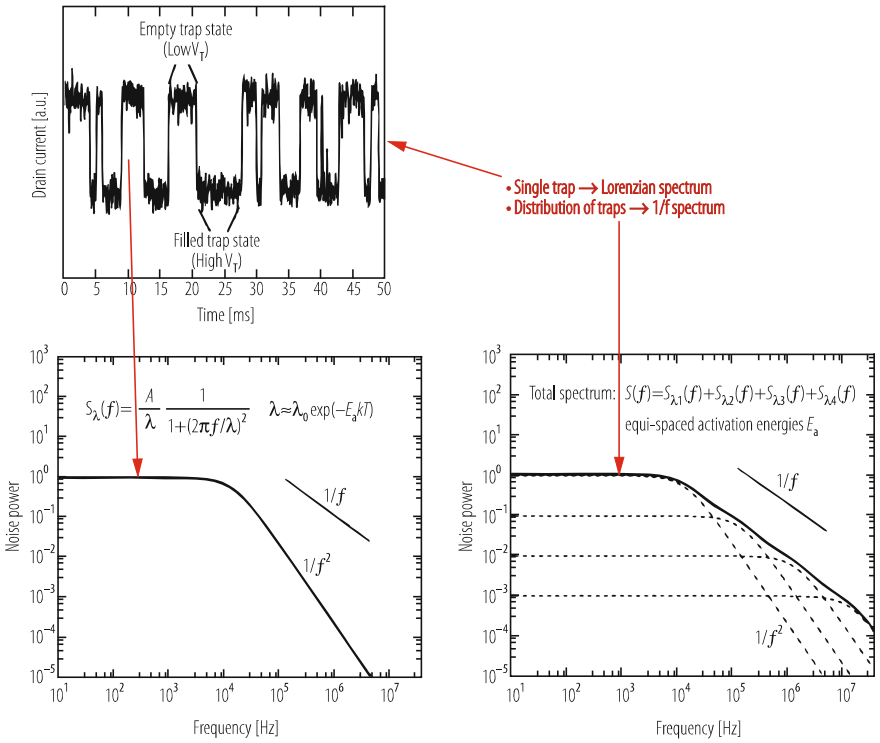


Fig. 10.11 Generation of random telegraph noise (left) and of 1/f noise (right) by trapping-detrapping. Adapted from Compagnoni et al. [17]

$1/|f|$ noise is one of the fractal processes, and its waveform preserves the same features independently of the time scale [16]. Another expression of this is independence of the measurement variance on the time scale of the measurement as long as the ratio of the high frequency and the low frequency cutoffs remains constant. As the bandpass moves along the frequency spectrum the spectral density integral (i.e., the measurement variance),

$$\sigma^2 \propto \int_{f_l}^{f_h} \frac{df}{f} = \ln\left(\frac{f_h}{f_l}\right) \tag{10.4}$$

remains constant for $f_h/f_l = const.$ In detector pulse processing it is well known that the contribution of $1/|f|$ noise to the equivalent noise charge (ENC) remains independent of the shaping time, as will be shown in the following Sect. 10.4.2. Physical mechanisms of $1/f$ noise are discussed in Ref. [21].

10.4 Noise in Charge Measurements

10.4.1 Sources of Noise in Charge Amplifiers

Principal noise sources in charge amplifiers and an equivalent diagram for calculation of the equivalent noise charge (ENC) are shown in Fig. 10.12.

Two elementary noise generators are included in the equivalent circuit, a series noise voltage generator representing the noise in the amplifying device, and a parallel noise current generator representing various noise sources not inherent to

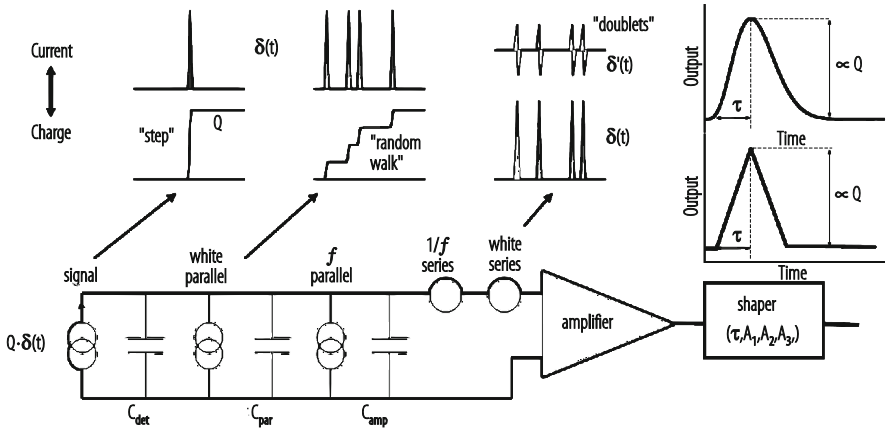


Fig. 10.12 An illustration of noise sources in charge amplifiers, as a Poisson sequence of elementary *current* pulses into the input capacitance, or *voltage* pulses on the capacitance

amplification (detector leakage current noise, parallel resistor noise, etc.). Both types of noise are assumed to have a white spectrum. Two forms of presentation in terms of a sequence of random pulses are shown, as charge (or voltage) at the input of the amplifier, and as a current injected into the input capacitance (comprised of the detector + amplifier parasitic capacitances). The presentation of the series noise in terms of a current into the detector input is the derivative of the charge (voltage) representation. The sequence of voltage impulses representing the amplifier series noise thus corresponds to an equivalent sequence of current doublets (derivatives of delta function) injected at the detector. The parallel noise is by its origin a current source in parallel with the detector, and it is presented by a sequence of impulses (delta functions). *It is this difference in the location of the two white noise sources with respect to the detector capacitance that makes their apparent noise spectra and their effect on the measurement quite different.* ENC due to the former is inversely proportional to the square root of the peaking time, and proportional to it due to the latter. The series $1/|f|$ noise contribution to ENC is independent of the peaking time, as indicated in Fig. 10.14. The $1/|f|$ noise due to a dissipative dielectric depends on the dielectric loss factor $\tan(\delta)$, as will be discussed in Sect. 10.4.7. It can be significant with detector-amplifier connections on glass fibre circuit boards.

10.4.2 Equivalent Noise Charge (ENC) Calculations

Calculation of equivalent noise charge (ENC) for a signal processing chain described by a weighting function $w(t)$ is summarized in the following.

The noise calculation is performed in the time domain by using Campbell's theorem Eq. (10.3), that is by superposition of effects of all random *current* impulses illustrated in Fig. 10.12. The weighting function is normalized to unity so that the definition of *ENC is the noise charge which produces an output of the same magnitude as an impulse signal of equal charge*. For calculation of ENC due to the series *voltage* noise, we will use the representation in terms of an equivalent current generator connected in parallel with the detector. This requires differentiation of the sequence of voltage impulses. Each resulting doublet $C_{in}\delta'(t)$ acting upon the weighting function $w(t)$ produces by convolution $C_{in}w'(t)$. The equivalent noise charge (ENC) is then given by,

$$ENC^2 = \frac{1}{2}e_n^2 C_{in}^2 I_1 + \pi C_{in}^2 A_f I_2 + \frac{1}{2}i_n^2 I_3, \tag{10.5}$$

where $\overline{e_n^2} = 4kTR_s$ is the physical (single-sided) noise *voltage* spectral density for series noise in V^2/Hz expressed in terms of an equivalent series noise resistance. The second term is due to the series $1/f$ noise. The $1/f$ noise physical spectral density is defined as A_f/f in [V^2/Hz]. The third term is due to the parallel noise, where $i_n^2 = 2qI_0 = 4kT/R_p$ is the physical noise *current* spectral density due to either a current or a resistance in parallel with the detector.

I_1, I_2, I_3 are the noise integrals for the *series* (*voltage*) white noise and the *1/f noise*, and for the *parallel* (*current*) noise, respectively. The integrals are derived in the time domain from Campbell's theorem Eq. (10.3) and expressed in the frequency domain using Parseval's theorem [9],

$$I_1 = \int_{-\infty}^{\infty} \left[w'(t)\right]^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(j\omega)|^2 \omega^2 d\omega = \frac{A_1}{\tau}, \tag{10.6}$$

$$I_2 = \int_{-\infty}^{\infty} \left[w^{(1/2)}(t)\right]^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(j\omega)|^2 \omega d\omega = A_2, \tag{10.7}$$

$$I_3 = \int_{-\infty}^{\infty} [w(t)]^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(j\omega)|^2 d\omega = A_3 \tau, \tag{10.8}$$

where τ is the time width parameter of the weighting function, either the peaking time of the function, or some characteristic time constant of the filter implemented in hardware (or software in case of digital filtering).

Noise contributions for both types of white noise due to various segments (piece-wise linear approximation) of the weighting function are shown in Fig. 10.13 (expressions for integrals I_1 and I_3). In these calculations, either the *impulse response* $h(t)$ of the system or the *weighting function* $w(t)$ (the mirror image of the impulse response) can be used for time-invariant systems. For time-variant (gated or switched) systems, only a weighting function describes the performance correctly, while an apparent impulse response (waveform at the output) is not correct and can be misleading. Steepest parts of the weighting function contribute most to ENC_s, as they correspond to larger bandwidth. Flat parts do not contribute anything. In contrast, ENC_p is largely proportional to the width of the weighting function where it has any significant value.

A bipolar weighting function, i.e., impulse response h(k), as shown in Fig. 10.13, with equal lobes would result in square root of two higher ENC than for a unipolar function (single lobe). If the amplitude of the second lobe is less than one half, its rms noise contribution becomes small (<12%).

The half-order integral I_2 for 1/*f* noise is not amenable to such a simple interpretation, and it will be discussed in Sect. 10.4.6.

Equations (10.6, 10.7, and 10.8) provide an insight into the general behaviour of signal processing systems with respect to noise. The ENC due to the series white amplifier noise is proportional to the slope (~1/t) of the weighting function and therefore proportional to the bandwidth of the system. The ENC due to parallel white noise is proportional to the width of the weighting function and therefore to the overall integration time. If the weighting function form remains constant the ENC due to *1/f* noise is independent of the width of the weighting function, since the ratio of the high frequency cutoff and low frequency cutoff remains constant,

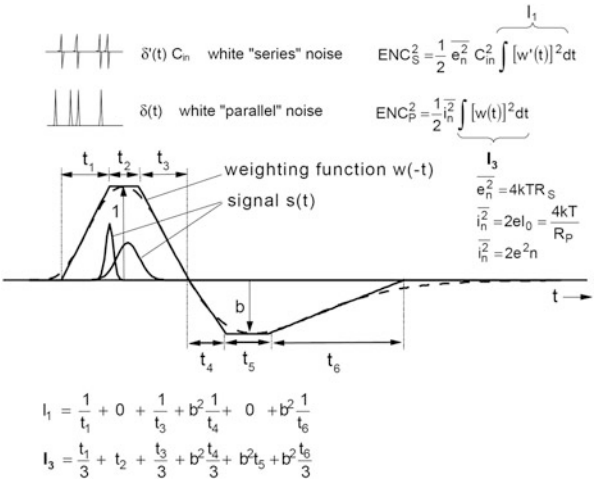


Fig. 10.13 An illustration of noise contributions for both locations of white noise sources (series and parallel) due to various segments of the weighting function. Such a piece-wise linear approximation of the weighting function provides an estimate of the noise within a few percent of accurately computed integrals I_1 *and* I_2

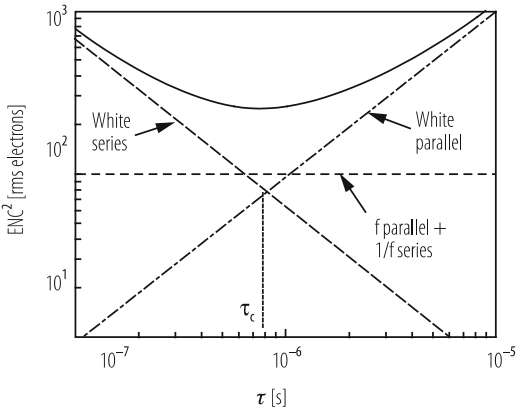


Fig. 10.14 An example of general behaviour of equivalent noise charge (ENC) as a function of the width parameter τ of the weighting function. 1/*f* noise raises the noise minimum but does not affect its position on the time scale

Eq. (10.4). This is illustrated in Fig. 10.14. From Eqs. (10.5, 10.6, and 10.8) the optimum width parameter of the weighting function is given by,

$$\tau_{opt} = \left(R_s R_p\right)^{1/2} C_{in} \left(A_1 \Big/ A_3\right)^{1/2}, \tag{10.9}$$

and it is not affected by 1/*f* noise.

10.4.3 Weighting Function

The concept of weighting function is very useful for time domain noise analysis of time variant, sampled and switched systems. The role of "pulse shaping," "signal filtering," or "signal processing" is to minimize the measurement error with respect to the noise, various baseline offsets and fluctuations, and at high counting rates to minimize the effects of pulse overlap or pileup. The term "pulse shaping" implies that the amplifier-filter system is time invariant. In such a system the system parameters do not vary during the measurement and a single measurement of amplitude or time is performed. Such a system is described completely by its impulse response.

In signal filtering, we also use time-variant methods, such as capacitor switching and correlated multiple sampling of the signal. The filtering properties of a time-variant system are described by its weighting function $w(t)$. *The weighting function describes the contribution that a noise impulse, occurring at time t_i, makes at the measurement t_m*, as illustrated in Fig. 10.15. It is essentially a measure of the memory of noise impulses (or any other signals) occurring before the observation time t_m. As shown, *the weighting function for time-invariant systems is simply a mirror image in time of the impulse response, with its origin displaced to t_m*. For a time-variant system, the impulse response (output waveform) is generally quite different from its weighting function. In some cases time-invariant and time-variant processing could be devised to produce the same result, i.e., both methods will be described by the same weighting function, while their implementation will be quite different. The noise-filtering properties of any weighting function for detector signal processing can be most easily determined by the time domain analysis technique shown in Fig. 10.13. The time domain analysis method based on Campbell's theorem was first introduced by Wilson [22], and subsequently elaborated in Refs. [9, 23, 24].

A composite weighting function for multiple correlated sampling is obtained by superposition of weighting functions for individual samples. This is illustrated

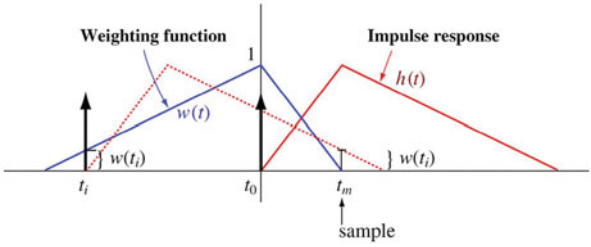


Fig. 10.15 An illustration of the weighting function $w(t)$ corresponding to impulse response $h(t)$. A unit noise impulse at t_i contributes $w(t_i)$ at the time t_m of the peak of response to the signal impulse at t_0. The weighting function in this case of a simple time-invariant filter is a mirror image of the impulse response delayed by the sample (i. e., measurement) time t_m

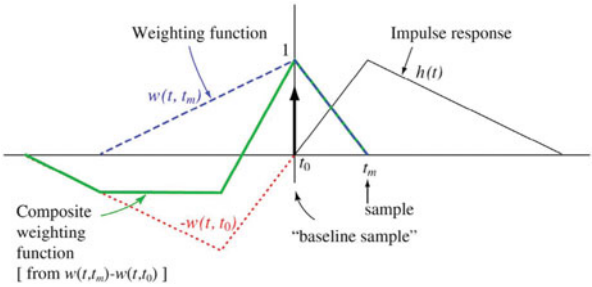


Fig. 10.16 Composite weighting function for *correlated double sampling* (CDS). Time-invariant filtering ("pulse shaping") described by an impulse response $h(t)$ is assumed prior to sampling

for *correlated double sampling* (*CDS*), a technique commonly used for readout of CCD's and large pixel arrays, Fig. 10.16. Single sample processing is described by a symmetrical triangular impulse response approximating single RC differentiation and one or two RC integrations. The single sample weighting function with respect to the sampling time at t_m is shown (dashed), and it is a mirror image of the impulse response. It is assumed that a (delta function) signal of interest will arrive at time t_{0+}, and produce a response described by the impulse response. In double correlated sampling another sample is taken at t_0, just before the arrival of the signal. This sample, sometimes called "baseline sample", is subtracted from the "signal or measurement sample". The weighting function for the baseline sample is shown inverted and earlier in time by t_m-t_0.

The composite weighting function (thicker solid line) is bipolar and it has area balance. This is another way of saying that CDS has zero dc response and that it attenuates (but does not eliminate) baseline fluctuations at low frequencies. The ENC can be easily calculated from such a composite weighting function using the technique for time domain noise analysis shown in Fig. 10.13. A noise analysis of such a case in the frequency domain and without the use of a composite weighting function is considerably more time consuming.

10.4.4 Simple ENC Calculation for Series White Noise

Following on the discussion in Sect. 10.4.2 and referring to Fig. 10.17, a simple relation for the equivalent noise charge (ENC_s) due to *series white noise* follows,

$$ENC_s^2 = (1/2)\, e_n^2 C_{in}^2 I_1, \tag{10.10}$$

where

$$I_1 = \int_{-\infty}^{\infty} \left[w'(t)\right]^2 dt = {}^{2}\Big/_{t_m} = {}^{A_1}\Big/_{t_m}, \tag{10.11}$$

$$ENC_s = {}^{e_n C_{in}}\Big/_{t_m^{1/2}}. \tag{10.12}$$

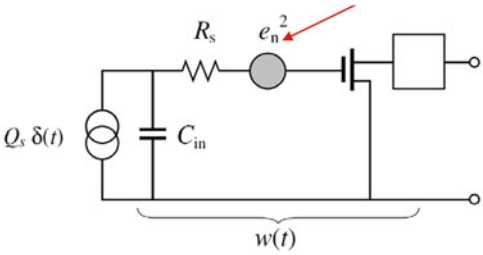


Fig. 10.17 Simplified equivalent circuit for calculation of equivalent noise charge (ENC) due to amplifier series white noise

It requires the knowledge of three parameters: noise spectral density e_n, total input capacitance (detector + amplifier) C_{in}, and peaking time t_m of the triangle approximating the weighting function. Such an approximation is useful for noise estimation, since the series noise coefficient for a fifth order semi-Gaussian weighting function with equal peaking time, $A_1 = 2.2$, differs by only ~10% from $A_1 = 2$ for the triangular function. In a preamplifier design, the expected e_n can be determined from the operating conditions (current and transconductance) of the first transistor, or from a more complete equivalent circuit of the input transistor shown in Sect. 10.5.2. *A primary objective of low noise amplifier design is to make the noise contributions of all other circuit components negligible compared to the input transistor.* Eq. (10.12) describes simply also the noise charge slope with respect to detector capacitance (in pF),

$$\partial\left(ENC_s/q_e\right)/\partial C = e_n/t_m^{1/2}\ \left[\text{rms electrons/pf}\right] \tag{10.13}$$

10.4.5 Simple ENC Calculation for Parallel White Noise

From Sect. 10.4.2 and Figs. 10.13 and 10.18 simple relations follow for ENC_p due to parallel shot noise and resistor (thermal) noise. The gated integrator case, where the weighting function equals unity for the duration of the gate, illustrates that the ENC_p for shot noise is simply the square root of the variance of a Poisson sequence of impulses counted for a time t_G .

$$ENC_p = (q_e I_0 t_G)^{1/2} = \left(q_e^2 n t_G\right)^{1/2} = q_e (n t_G)^{1/2} = q_e n_G^{1/2}. \tag{10.14}$$

By Campbell's theorem the contribution of each impulse to the variance is determined by the weighting function, and for a given weighting function the parallel noise integral I_3 has to be determined. For an approximation by a triangle with a peaking time t_m, $I_3 = (2/3)t_m$. The parallel noise contribution for the triangular weighting function is the same as for gated integration one third as wide.

The contribution by the parallel resistor thermal noise can be compared simply to the shot noise by the "*50 mV rule*": a dc current I_0 causing a voltage difference of

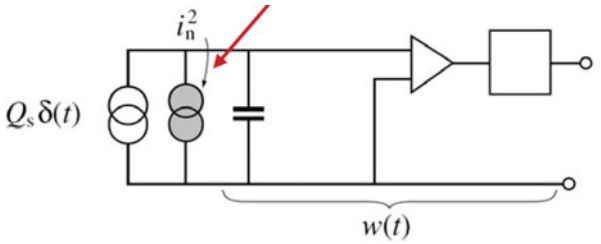


Fig. 10.18 Simplified equivalent circuit for calculation of equivalent noise charge (ENC) due to detector and amplifier parallel white noise (bias or feedback resistance, detector leakage current, tunneling gate current in MOS, base current bipolar junction transistor)

~50 mV on a resistor R_p contributes a shot noise equal to the thermal noise of that resistor at room temperature; from $4kT/R_p = 2q_e\, I_0$: $R_p\, I_0 = 2kT/q_e$.

10.4.6 Calculation and Estimation of ENC for 1/f Noise

1/*f* noise becomes a limiting factor in many physical measurements. We can imagine reducing the series white noise in charge measurements to a very low level by continuing to increase the measurement (integration) time τ, provided the parallel (leakage or dark current) noise is very low. We would eventually reach the "noise floor" due to the 1/*f* noise. Once the 1/*f* noise spectral density is determined experimentally and defined by the parameter A_f in [V^2] as in Eq. (10.5), ENC can be calculated by the integral I_2, Eq. (10.7). In the time domain this is an integral of a fractional-order (half-order) derivative squared of the weighting function (a mathematical operation which cannot be called "trivial" before one learns how to do it, and it can be considered "tedious" at best). In the frequency domain the calculation is somewhat easier for time-invariant systems, but for time-variant systems defining the transfer function $H(j\omega)$ is more difficult and less intuitive than determining the weighting function.

We illustrate this here on the example of a commonly used weighting function of trapezoidal form as shown in Fig. 10.19. There are many different hardware implementations of this function in different applications. Time-invariant versions have used delay line clipping and higher order RC prefilters. Gated integrator and higher order prefilters have been used in several applications, starting with germanium gamma-ray detectors [25]. This function is widely used with CCDs in astronomy, implemented by correlated double sampling and dual-ramp integration.

We define the trapezoidal weighting function by the width of the ramp τ_p and the flat top as a fraction of the ramp, $\Delta\tau_p$. The equivalent noise charge for 1/*f* noise is then,

$$ENC_f^2 = \pi C_{in}^2 A_f A_2, \tag{10.15}$$

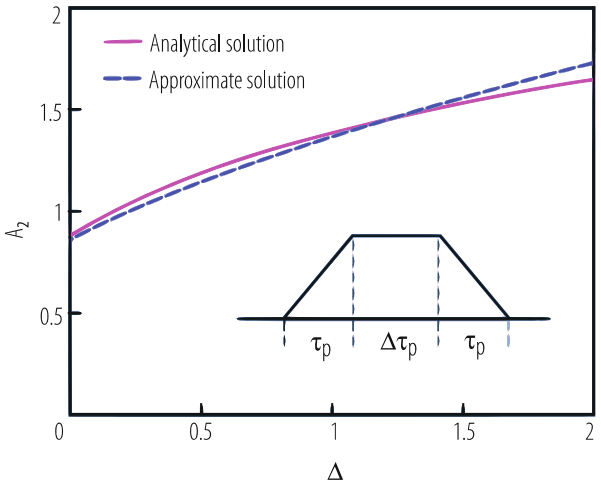


Fig. 10.19 1/*f* noise coefficient A_2, Eq. (10.7), for trapezoidal weighting function. *Solid line*: analytical solution, Eq. (10.18); *dashed*: approximate solution, Eq. (10.22)

where

$$A_2 = \int_{-\infty}^{\infty} \left[w^{(1/2)}(t) \right]^2 dt. \tag{10.16}$$

The half order derivative of the weighting function, $w^{(1/2)}(t)$, is obtained by convolution of $w'(t)$ with the elementary pulse $U(t)/t^{1/2}$ for 1/*f* noise shown in Fig. 10.10d,

$$w^{(1/2)}(t) = \left(1/\sqrt{\pi t} \right) * w'(t), \text{ for } t \geq 0+ \tag{10.17}$$

Using *Mathematica,* and after some manipulation, the result for A_2 is, [26],

$$A_2 = \frac{1}{\pi} \left[\Delta^2 \ln \Delta + (2+\Delta)^2 \ln (2+\Delta) - 2(1+\Delta)^2 \ln (1+\Delta) \right]. \tag{10.18}$$

The coefficient A_2 vs the flat top Δ of the trapezoidal weighting function is plotted in Fig. 10.19.

The effect of the series 1/*f* noise is lowest for a triangular weighting function, $\Delta = 0$, which is also the case for the series white noise, Eqs. (10.10, 10.11, and 10.12). As the flat top is made longer, A_2 increases, since such a trapezoidal function has a higher ratio of its cutoff frequencies, which results in integrating a wider band of the *1/f* noise spectrum, Eq. (10.4).

An almost identical result for A_2 has been obtained by a calculation in the frequency domain and described in a study of CCD noise performance [27].

Since such exact calculations of ENC_f for any weighting function can be time consuming, we emphasize here a simple estimation method, which provides results sufficiently close to the exact calculations for most purposes. It has been pointed out by Gatti et al. [28] that the three integrals in Eqs. (10.6, 10.7, and 10.8) have to satisfy the Cauchy-Schwartz inequality,

$$\int_{-\infty}^{\infty}\left[w^{(1/2)}(t)\right]^2 dt \le \left\{\int_{-\infty}^{\infty}\left[w'(t)\right]^2 dt \bullet \int_{-\infty}^{\infty}[w(t)]^2 dt\right\}, \tag{10.19}$$

that is,

$$A_2 \le (A_1 A_2)^{1/2}. \tag{10.20}$$

Thus there is an upper limit to A_2 in relation to A_1 and A_3 which are easily calculated from Fig. 10.13, or Eqs. (10.6, 10.7, and 10.8). A study of the most commonly used weighting functions, [28], reveals that $A_2/(A_1A_3)^{1/2}$ falls between 0.64 and 0.87, a spread of less than ±8% in the calculation of rms noise, so that for the estimation of 1/*f* noise the following approximate relation can be used,

$$A_2 \approx 0.75(A_1 A_3)^{1/2}. \tag{10.21}$$

For the trapezoidal weighting function in Fig. 10.19, $A_1 = 2$ and $A_3 = \Delta + 2/3$, and the approximation for this case is,

$$A_2 \approx 0.75\left[2\left(\Delta + {}^{2}\!/_{3}\right)\right]^{1/2}. \tag{10.22}$$

Figure 10.19 shows that this approximation is within a few percent of the exact analytical solution, Eq. (10.18).

In any noise analysis of charge amplifiers one will have already calculated, or otherwise determined the values of A_1 and A_3, so that the information about the filtering (pulse shaping) effect on the series and parallel white noise will readily also provide an estimate of the 1/*f* noise,

$$ENC_f^2 = \pi C_{in}^2 A_f A_2 \approx \pi C_{in}^2 A_f \left(0.75\sqrt{A_1 A_3}\right). \tag{10.23}$$

It is interesting to note that for a Gaussian weighting function $A_2 = 1.00$, for a triangular weighting function 0.88, for a fourth order semi-Gaussian 1.02, for CR-RC 1.18.

A_f is a parameter resulting from a measured spectral density and it does not contain any specific information about the properties of the amplifying device unless other parameters are known.

For input transistor optimization a parameter which is to the first order independent of the device dimensions is more useful [29], $K_f = A_f\ C_{gs}$ [J]. This constant

ranges from 10^{-27} J for junction field-effect transistors (JFETs) to ~10^{-25} J for p-channel and ~ 10^{-24} J for n-channel MOS transistors.

For an accurate calculation of the noise charge for noise spectra departing from the three-term power-law representation ("white series voltage noise", "1/*f* series voltage noise", and "white parallel current noise"), circuit simulation and numerical calculation are the tools of choice to obtain accurate results. The discussion here was intended to provide some insight: *ENC_f depends only on the shape of the weighting function but not on the time scale.*

10.4.7 Noise in Dielectrics

Thermal fluctuations in dielectrics generate a noise which is quantitatively related to the parameters describing dielectric losses. This type of noise and its importance for detectors was first studied in [30] and then summarized in [31]. For a dielectric with low losses, the dissipation factor or the loss factor D (equal to the imaginary part ε "of the permittivity $\varepsilon = \varepsilon' + j\varepsilon''$) is independent of frequency in the range of interest for particle and photon detectors (~10^4 to 10^8 Hz). It can be defined as $D = G(\omega)/(\omega C_{diel})$, where $G(\omega)$ and C_{diel} are the loss conductance and the capacitance of the dielectric as measured on an impedance bridge at an angular frequency ω. According to the fluctuation-dissipation theorem [32, 33], and using the Johnson-Nyquist formula for thermal noise, a dissipative dielectric generates a noise current with a spectral density,

$$i_n^2 = 4kTG(\omega) = 4kTD\omega C_{diel}. \tag{10.24}$$

The equivalent noise charge ENC_{diel} due to dielectric noise can be calculated using Eq. (10.7),

$$ENC_{diel}^2 = 2kTDC_{diel}A_2. \tag{10.25}$$

Following on the discussion in the previous section we assume here $A_2 = 1.2$,

$$ENC_{diel}^2 = 2.4kTDC_{diel}. \tag{10.26}$$

We note that the spectral density (Eq. 10.24) upon integration on the input capacitance becomes 1/*f*, and therefore the equivalent noise charge due to dielectric noise is independent of the width (the time scale) of the weighting function.

The noise from lossy dielectrics may pose in some detectors a lower limit to total noise. If a lossy dielectric contributes 1 pF, such as a glass fibre board with $D \approx 2{\cdot}10^{-2}$, this alone would present a lower limit of $ENC_{diel} \approx 86$ rms electrons. Best dielectrics (e.g. Teflon, polystyrene, quartz) have $D \approx 5{\cdot}10^{-5}$, which results in ~5 rms electrons. This noise contribution to the charge measurement can be reduced only by reducing C_{diel} (and/or the temperature).

10.5 Gain Mechanisms and Noise in Transistors

10.5.1 Gain Mechanism, Electron Transit Time, Unity-Gain Frequency

The charge control concept as the basis for the gain mechanism in all three-terminal amplifying devices (transistors) was discussed in Ref. [9]. The "control charge" Q_c is illustrated in Fig. 10.22.

The relation among the control capacitance C_{gs}, the transconductance g_m, the electron transit time τ_e and the unity gain frequency f_T is summarized by,

$$\begin{aligned} &\Delta Q_c/\Delta I_d = \tau_e = C_{gs}\Delta V_{gs}/\Delta I_d = C_{gs}/g_m \rightarrow C_{gs} = g_m\tau_e, \\ &f_T \approx {}^{1}\Big/_{(2\pi\tau_e)} = \left({}^{1}\Big/_{2\pi}\right)\left({}^{g_m}\Big/_{C_{gs}}\right). \end{aligned} \tag{10.27}$$

NMOS transistors in submicron range (channel length below ~0.25 microns) will have a unity gain frequency in the range 10 to 100 GHz when operated in strong inversion. These same devices will be operated in weak inversion to maximize the transconductance/current ratio and the power dissipation in detectors with large numbers of channels (pixels or strips). This means reduced transconductance with almost the same gate capacitance resulting in a unity gain frequency in the range of 1 GHz or less. This affects the speed of response and the stability considerations in the design of feedback amplifiers.

Equation (10.27) describe only a simplified basic relation among intrinsic device parameters. A very extensive treatment of charge control concepts for CMOS transistors including parasitic parameters and device operating conditions is given in Ref. [7].

10.5.2 Noise Sources in MOS Transistor

A brief overview of white noise sources in an NMOS transistor normalized to the intrinsic channel series noise resistance γ/g_{ms} is illustrated in Fig. 10.20 and Eq. (10.28) with γ typically in the range 0.5 to 1.0. The second term in Eq. (10.28) is the gate induced noise contribution [34]. The coefficient $\delta/5\gamma$ depends on the bias conditions. For estimation purposes $(\delta/5\gamma) \sim 1/3$. With capacitive sources, such as most radiation detectors, this term is usually negligible. In particular, at operating conditions to minimize the power in the input transistor, the optimum ratio C_{gs}/C_{in} is small. The contributions by the gate resistance and substrate resistance can be made

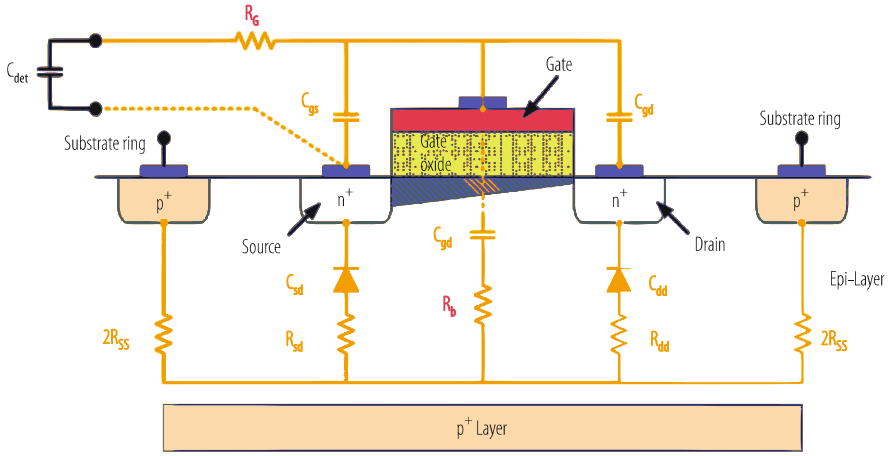


Fig. 10.20 An illustration of parasitic resistive noise sources in an NMOS transistor in addition to the channel noise γ/g_{ms}

small by the device design. The equivalent series noise resistance of the NMOS transistor can be summarized referring to the notation in Fig. 10.20 as,

$$\frac{R_{eq}}{\gamma/g_{ms}} = 1 + \frac{\delta}{5\gamma}\left(\frac{C_{gs}}{C_{in}}\right)^2 + \frac{1}{\gamma}\left(R_g g_{ms}\right) + \left(R_b g_{mb}\right)\frac{g_{mb}}{g_{ms}} + \frac{1}{\gamma}\left(R_b g_{ms}\right)\left(\frac{C_{gb}}{C_{in}}\right)^2. \tag{10.28}$$

10.5.3 Charge Transfer from Detector to Transistor: Capacitance Matching

In all cases where the amplifier is connected directly to the detector via a resistive conductor the charge produced by ionization is distributed among the detector capacitance, amplifier capacitance and any stray capacitance according to the ratio of capacitances, Fig. 10.21. Due to this, only a fraction of the charge of interest (the signal) arrives where it matters—that is to the conduction channel of the input transistor where it controls the drain (collector) current, as illustrated in Fig. 10.22. (An exception to this is if the two capacitors are connected by an inductor in which case the charge is transferred periodically between the two capacitors.) In case of a CCD the ionization charge is moved peristaltically in a potential well formed and driven by appropriate clock voltages applied to the gate electrodes. The charge shifted a few hundred (or thousand) times arrives at the collection electrode ("floating diode") which is connected to a source follower. In the CCD the charge arriving at the collection electrode is the original charge packet produced

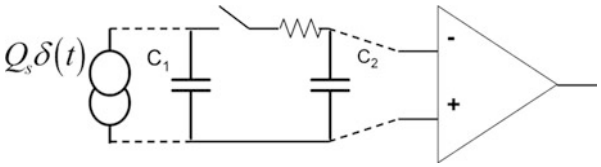


Fig. 10.21 Sharing of the induced ("collected") charge between the detector and amplifier capacitance

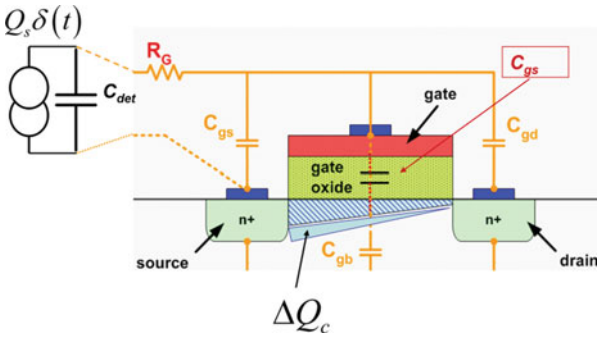


Fig. 10.22 An illustration of the *control charge* with respect to the *control capacitance* C_{gs}: $\Delta Q_c = Q_s/[1 + (C_{det} + C_{gsp} + C_{ds})/C_{gs}]$

by ionization except for a few electrons lost to trapping. The charge transport in a conductor is by a small displacement of a large number of free electrons. The CCD principle allows multiple measurements on the same charge packet as described in Ref. [35].

Optimization of the signal to noise ratio requires appropriate matching of the transistor active capacitance (which controls the current) to all other capacitances connected to the input—a subject addressed in some detail in Ref. [10].

The charge control concept expressed by Eq. (10.27) is at the basis of the gain mechanism in almost all electronic amplifying devices: it takes an increment of charge, ΔQ_c in Fig. 10.22, to cause a steady state change of the current in the conducting channel. From Eqs. (10.12 and 10.27) we can determine the lower limit for the charge sensitivity due to the series noise in terms of the electron transit time in the conducting channel. Scaling the device width (with no power limitation) to achieve the best signal-to-noise ratio requires that the control electrode capacitance equal (match) the detector capacitance, $C_{gs} = C_{det}$. With this and Eq. (10.27), the lowest noise for an electronic amplifier that could be achieved under ideal circumstances is given by,

$$ENC_{s\ opt} \cong 2\sqrt{2}\sqrt{kTC_{gs}}\sqrt{\tau_e / t_m}, \tag{10.29}$$

where t_m is the weighting function zero-to-peak time, as in Fig. 10.13, also referred to as the "integration time" or the "measurement time".

It is assumed here for simplicity that the equivalent series noise resistance is equal to the inverse of the device transconductance, i.e., $\gamma \sim 1$. While Eq. (10.29) is useful for estimation purposes and for establishing a lower limit for the amplifier series noise, the electron transit time is rarely used directly for noise calculations.

Accurate noise calculations usually rely upon the measured or known noise spectral density for a particular device and have to include parasitic capacitances and resistances, as indicated in Fig. 10.20.

10.6 Weighting Function Realizations: Time Invariant and Time Variant

10.6.1 Time-Invariant Signal Processing

A low order asymmetrical function (dashed), shown in Fig. 10.23, results in higher series noise (due to a steeper rise, see Fig. 10.13). A nearly symmetrical function requires a higher order signal processing chain, as shown.

A time-invariant circuit realization of a nearly symmetrical weighting function is described in Ref. [36].

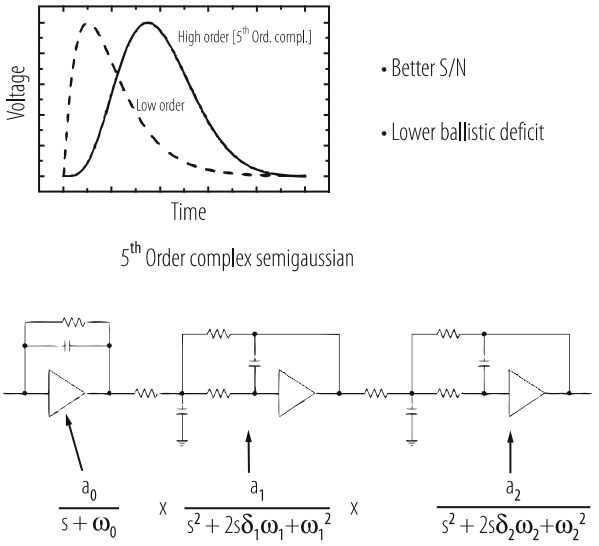


Fig. 10.23 A typical time-invariant signal processing chain resulting in a fifth order pseudo-Gaussian impulse response (weighting function), where s is the Laplace transform variable. This function results in lower noise coefficients A_1, A_2, A_3, Eqs. (10.6, 10.7, and 10.8), 2.2, 1.05, and 0.78, respectively, compared to a low-order CR-RC function *of equal width at the base*. See Ref. [10] for design considerations and other implementations

10.6.2 Uncorrelated Sampling and Digital Filtering

Sampled data digital signal processing has become prevalent in detector systems for gamma-ray and x-ray spectroscopy, for time projection chambers and in various forms in particle physics. One of the advantages is that it provides flexibility in the realization of mathematically optimal weighting functions. Optimal signal processing cannot be achieved without some analogue functions. An anti-aliasing filter is an essential part of the system. Its function is to limit the bandwidth prior to sampling so as to satisfy the Nyquist-Shannon sampling criterion: the bandwidth at the output of this filter must be no more than one half of the sampling frequency (the "Nyquist limit"). If this is not satisfied, the noise at frequencies higher than the Nyquist limit is shifted in frequency, i.e., ("aliased") by undersampling, to the frequencies below this limit. *The resulting loss in S/N due to aliasing cannot be recovered by any subsequent processing.* The role of digital filtering is to create optimized weighting functions in spectroscopic systems, and to enable an optimal particle track measurement in tracking systems. In spite of the power of digital processing, it is most efficient to cancel any long tails in the detector-preamplifier response by analogue means. If the tail cancellation is performed digitally, much larger numbers of samples have to be processed (deconvolved) for each event. For asynchronous (uncorrelated) sampling in semiconductor detectors for gamma-ray and x-ray spectroscopy see Refs. [37, 38]. In such systems optimum weighting functions are of trapezoidal form as in Fig. 10.19, where the flat top allows uniform weighting for a variable charge collection time.

10.6.3 Correlated Sampling

Correlated double sampling (CDS) is being used with many detectors in various implementations and under different names. One of these is known as "baseline subtraction"—taking a sample prior to the arrival of the usually unipolar signal (a delayed signal or with the arrival time known). This case is illustrated in Fig. 10.16, and it results in a bipolar weighting function, which defines quantitatively the effect of CDS on the noise, as discussed in Sects. 10.4.2 and 10.4.3. Correlated double sampling is an essential part of CCD signal processing in astronomy.

Signal processing by *multiple correlated (synchronous) sampling* has been used for noble liquid calorimeters, such as the liquid argon electromagnetic calorimeter in the ATLAS experiment, Refs. [39, 40].

10.7 Equipartition and *kTC* Noise

Integration of the power spectrum (spectral density) arising on a capacitance from the thermal noise current, $i_n^2 = 4kT/R$, of the resistor results in the *total fluctuation* of charge (and voltage), which is *independent of the value of the resistance R*. The bandwidth (equivalent to an abrupt cutoff) of the RC circuit in Fig. 10.24 is $1/(4RC)$, and the total fluctuation in voltage and charge is

$$\begin{aligned} \sigma_v^2 &= kT/C, \\ \sigma_q^2 &= kTC. \end{aligned} \tag{10.30}$$

The resistance (with the capacitance C) determines the bandwidth of the noise but not its magnitude. The *kTC* noise at 300 K is quite high even on small capacitances, as shown in Table 10.1. Most of this noise does not affect the measurement in systems where filtering and a very high parallel (feedback) resistance is used following a preamplifier, in other words where the noise corner time constant, Eq. (10.9), is much longer than the width of the filter weighting function. Such a system responds only to the portion of the spectrum where the spectral density is very low. *An example*: high resolution x-ray spectrometry with silicon detectors, where time-invariant or digital filtering is usually used. In contrast, when the measurement is performed by taking a sample directly on the detector capacitance and the filtering is not possible, the full *kTC* noise is included in the measurement, and it can be reduced only by correlated double sampling (CDS)—if applicable, as discussed in Fig. 10.27. CDS is just another way of excluding the noise at low frequencies from the measurement.

From the above discussion, which is based on circuit analysis, one is led to conclude that the *kTC* noise arises from the resistance, and yet its magnitude is independent of the value of the resistance. One may also be led to conclude that an "ideal" capacitance would have no noise. This is contradicted by the equipartition theorem which makes no direct assumption about the resistance. The equipartition

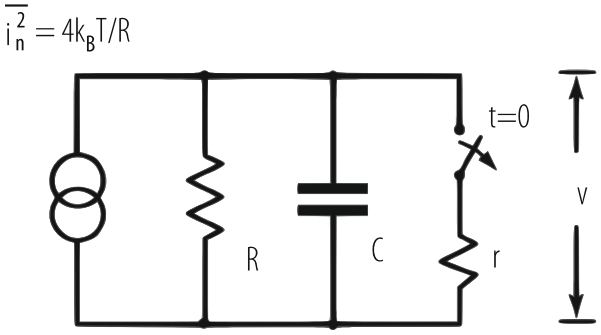


Fig. 10.24 A simple circuit for calculation of *kTC* noise

Table 10.1 Charge and voltage total fluctuation vs capacitance ($T = 300$ K)

Capacitance	Charge fluctuation	Voltage fluctuation
C [F]	$(kTC)^{1/2}/q_e$ [rms e]	$(kT/C)^{1/2}$ [μV]
1a	0.4	$6.4{\cdot}10^4$
10a	1.26	$2.0{\cdot}10^4$
100a	4	$6.4{\cdot}10^3$
1f	$1.26{\cdot}10$	$2.0{\cdot}10^3$
10f	$4.0{\cdot}10$	$6.4{\cdot}10^2$
100f	$1.26{\cdot}10^2$	$2.0{\cdot}10^2$
1p	$4.0{\cdot}10^2$	64
10p	$1.26{\cdot}10^3$	20
100p	$4{\cdot}10^3$	6.4

theorem states, that for a system in thermal equilibrium, the fluctuation energy per degree of freedom is *kT/2*. "*Per degree of freedom*" applies to any variable by which the energy of an energy storage object, or energy storage mode, can be defined. Thus for a capacitance, $C\langle v^2\rangle = kT/2$, from which Eq. (10.30) follows. So the statistical mechanics gives the same result for the total fluctuation but without any details about the dissipative components and the noise spectrum. A practical consequence is that as a capacitor becomes closer to an ideal one, the noise spectrum shifts toward zero frequency, while the total fluctuation remains constant.

We add here parenthetically that in a resonant system (an inductance- capacitance circuit), where there are two degrees of freedom, the noise spectrum is concentrated around the resonant frequency, while the integral of the power spectral density (total charge fluctuation) on the capacitance equals *kTC,* and the total current fluctuation (variance) in the inductance equals *kT/L*.

The above considerations apply also to analogous mechanical systems.

Transient behaviour of *kTC* noise is of great interest for switched capacitance circuits and for pixel detectors, where pixels are read out directly without filtering, by being sampled in a matrix arrangement, either before or after simple amplifiers (source followers, or three transistor circuits as in, Figs. 10.32, 10.33, and 10.34).

Transient behaviour of noise on a capacitance after switching the resistance or capacitance can best be studied by applying Campbell's theorem, as shown in Figs. 10.25 and 10.26. In this case the integration of the variance has to be carried out from the time when the switching takes place (zero) to the time t when the observation is made,

$$\sigma_v^2 = \left(\frac{1}{2}\right) 4kT \frac{1}{R} \int_0^t h^2(u)du = \frac{kT}{C}\left(1 - \mathrm{e}^{-2t/\tau}\right). \tag{10.31}$$

The oscillogram shows build-up of noise after switching a white noise source onto an RC circuit, $h(t) = (1/C)\mathrm{e}^{-t/\tau}$. The time constant τ equals the product of the capacitance and the resistance after the switching. Such a build up occurs after

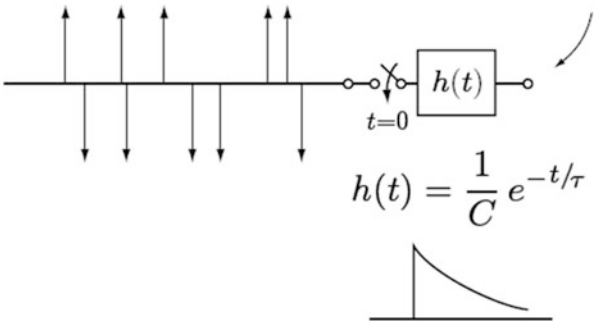


Fig. 10.25 Model for calculation of the build-up of *kTC* noise on an RC circuit

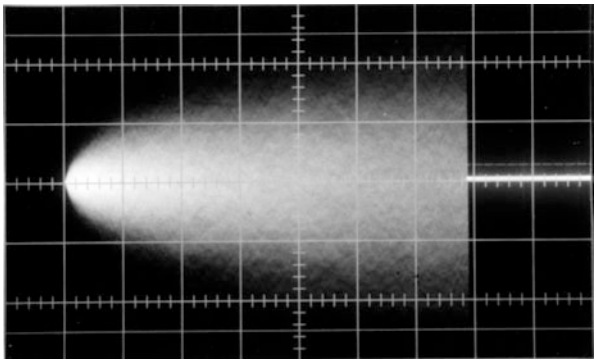


Fig. 10.26 Build-up of kTC noise on an RC circuit

a reset switch across a capacitor is opened and a much higher value of resistance appears in parallel with the capacitance. This is illustrated in Fig. 10.27.

Figure 10.27 illustrates what happens with *kTC* noise in active pixel sensors and CCDs. While the reset switch is closed, the *kTC* noise extends to very high frequencies corresponding to the very short time constant $r_{ON}C$. When the switch is "opened" the time constant increases by many orders of magnitude. A value of the "*old*" *kTC* noise is stored on the capacitance, and it decays very slowly with this very long time constant, CR_{OFF}. It is this sample of the "*old*" wide bandwidth noise that is often referred to as the "*reset noise*", even though its origin is not in the reset action. During the same time after switching, the "*new*" *kTC* noise builds up also very slowly, but faster than the stored value decays, since the rms noise build-up proceeds as $(1 - \exp[-2\, t/CR_{OFF}])^{1/2}$, Eq. (10.31) and Fig. 10.27. From this illustration one can see the conditions under which correlated double sampling may reduce significantly the *kTC* noise: *Sample 1 may be taken any time between opening of the reset switch and the arrival of the signal. Sample 2 may be taken any time after the arrival of the signal but before the "new" kTC noise has built up.*

Analysis of the effects of the *kTC* noise in some cases is not straightforward. Here are some general guidelines:

- *Fluctuation-dissipation theorem* with Johnson-Nyquist expression for thermal noise is essential for calculation of noise spectra and for detailed information on noise sources based on circuit analysis.

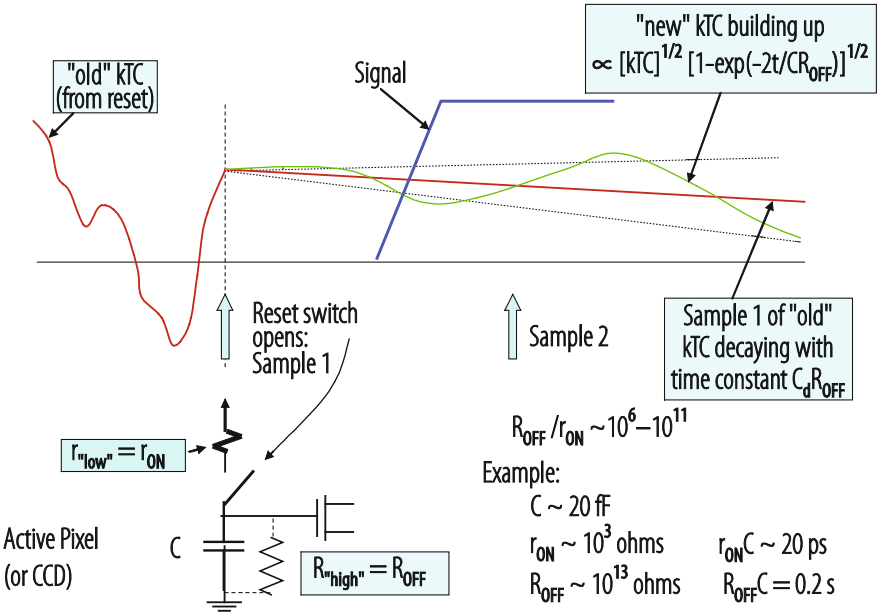


Fig. 10.27 Transient behaviour of *kTC* noise caused by the reset action of the sense node in CCDs and in pixel detectors with matrix readout (e.g., hybrid CMOS detectors)

- *Equipartition theorem* provides no detailed information on the noise spectra, but provides a *check on the integrals* of noise spectra (the total fluctuation).
- *Transient behaviour* of noise in switched capacitor circuits and matrix readout pixel arrays is best understood by means of *Campbell's theorem*, which provides *noise variance vs time*, as shown in Figs. 10.25, 10.26, and 10.27 and by Eq. (10.31). The knowledge of the dissipative component (resistance) is necessary for the transient analysis.
- A charge reset and transfer by a switch result in *kTC* independently of the switch ON resistance. This noise can be subtracted only if the *first sample* in the CDS is taken *before the signal*.
- Transfer (i.e., direct transport) of charge without switching (as in a CCD) does not result in *kTC* noise. *Reset* of the sense node *does*.

A frequently asked question: Can the total charge fluctuation (variance) on a capacitance be reduced below *kTC*? Yes, by "electronic cooling", where the apparent noise temperature of the resistance in parallel with the capacitance is reduced by feedback, Sect. 10.8.1 and Ref. [45]. While this is useful in practice, a note should be made that a system with active elements (gain) cannot be considered as being in thermal equilibrium with the surrounding.

An important distinction between two classes of signal processing schemes should be emphasized: (1) when no filtering (band limiting) takes place before the measurement, the total *kTC* fluctuation will contribute fully to the charge

measurement, and usually plays a dominant role. In contrast, (2) with charge amplifiers followed by filtering, the contribution of this noise (then usually referred to as "parallel noise") can be made negligible in most cases, by avoiding the low frequency part of the spectrum (by using a short peaking time, Fig. 10.14). Correlated double sampling (CDS) is one form of filtering.

10.8 Some Basic Signal Processing and Detector Readout Circuits

10.8.1 Charge Amplifier Configuration

In the most basic charge amplifier feedback configuration only two transistors are essential to realize a complementary cascode, as shown in Fig. 10.28. The current sources in positive and negative supplies can be realized by resistors or by low noise transistor current sources. There is only one significant pole (C_0R_0) in the first order solution for the response of this circuit. Higher order poles are given by the unity gain frequency of the transistors used. The cascode alone is an "operational transconductance amplifier" (very high output impedance). With the follower amplifier $\times 1$ it becomes an operational amplifier.

Gain and input impedance relations for the feedback charge amplifier configuration are derived from Figs. 10.28 and 10.29. The frequency dependence of the open loop gain is inherent to a high gain single pole amplifier. It is described by two parameters, unity gain frequency $\omega_h = g_m/C_0$, and the gain "roll off" frequency (3 dB point) $\omega_l = 1/(R_0C_0)$. The dc gain is then $|G_0| = \omega_h/\omega_l = g_m/R_0$, where g_m is the transconductance of the input transistor, C_0 is the dominant pole capacitance and R_0 is the dominant pole resistance. Input impedance with capacitive feedback has two terms, a resistance $R_{in} = 1/\omega_h C_f$ in series with a capacitance C_f G_0. The resistance term R_{in}, in conjunction with the total input capacitance, determines the

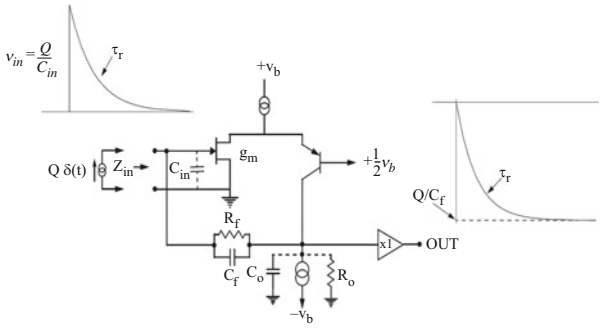


Fig. 10.28 Basic folded cascode charge amplifier feedback configuration

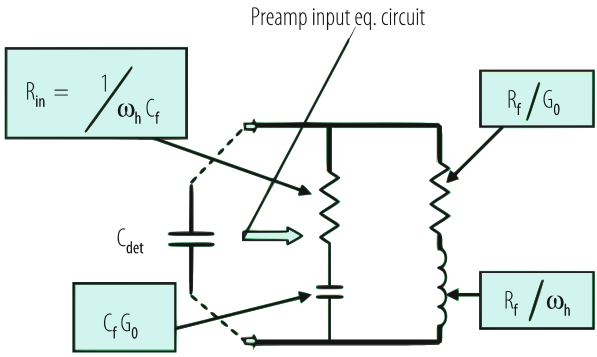


Fig. 10.29 Input equivalent circuit of feedback charge amplifier

rise time of the detector-amplifier. The rise time constant of the output voltage (i.e., the transfer of charge from the detector capacitance to the feedback capacitance) is $\tau_r = R_{in}C_{in} = (1/\omega_h)(C_{in}/C_f) = (C_0/g_m)(C_{in}/C_f)$, where $C_{in} = C_{det} + C_{ampl}$.

The resistive input impedance has a noise corresponding to the amplifier series noise resistance R_{seq}, and it appears as a resistance with a noise temperature, $T_{\text{eff}} = TR_{seq}/R_{in}$. For values of R_{in} higher than R_{seq}, the amplifier can be used as a termination for delay lines with a noise lower than that of a termination with a physical resistor Z_0 at temperature T.

The apparent noise temperature of the resistance Z_0 realized by the capacitance in feedback is $TR_{seq}/Z_{0,}$ and this is why it can be called "electronically cooled termination" or "electronically cooled damping", Ref. [45]. The resistance in parallel with the feedback capacitance adds two more terms to the input impedance of the preamplifier: inductance R_f /ω_h in series with a resistance R_f/G_0. It is important to note the condition to achieve an aperiodic ("damped") response of the feedback amplifier.

The feedback configuration allows the ultimate in noise performance because the parallel noise sources can be made negligible by using a transistor with a very low gate leakage current and a very high feedback resistance (megaohms to gigaohms). The feedback resistor can be avoided altogether by the use of optoelectronic feedback or a transistor switch to maintain amplifier voltages in the operating range. Signal integration is performed on the feedback capacitance C_f. The long tail can be cancelled in subsequent pulse shaping by a simple pole-zero cancellation circuit (not shown in the figure). Pulse shaping at the preamplifier by reducing R_L or R_f would result in increased noise from the thermal noise of these resistors. The object of the design is to avoid dissipative components at the detector-amplifier input and thus to make R_f as large as possible.

10.8.2 Cascaded Charge Amplifier Chain with Pole-Zero Cancellation

While the basic charge amplifier concept of a simple cascode circuit with a single dominant pole and capacitive feedback has not changed since the days of vacuum tube technology, the charge restoration techniques to control the operating point of the input transistor have evolved, particularly with the advent and widespread application of CMOS monolithic circuit technology. The dc feedback in charge amplifiers via a resistor (R_f in Fig. 10.28) has always been a problem in applications striving to achieve the ultimate in noise performance. Very high values in the gigaohm range are required in x-ray and gamma-ray spectroscopy, as this resistor injects a noise current inversely proportional to this resistance directly into the input node (i.e., "parallel noise", Sect. 10.4.5). In most applications the resistance values required are higher than the practical range of polysilicon resistors in CMOS technology. In some applications where the detector capacitance is very low it is best to avoid entirely any resistor and any continuous dc feedback. Various charge restoration techniques using switching or "reset" have been developed. An example is the CCD readout as shown in Fig. 10.35.

Continuous charge restoration is usually simpler to implement and better suited in many applications where its noise contribution is negligible, i.e., with higher detector capacitances and shorter weighting functions (peaking times). Most present gas, noble liquid and silicon particle detectors fall into that category. Additional considerations in the choice between switched and continuous charge restoration are the knowledge of the event arrival time and whether switching transients pose a problem. Continuous feedback in MOS technology is realized by a transistor with a long and narrow channel (L»W). The resistance of such a device will depend on the signal amplitude and on the detector leakage resulting in a nonlinear response. An elegant solution for accurate nonlinearity compensation is shown in Fig. 10.30. The transistor-capacitor network between the two amplifiers is an exact replica of the feedback network but increased in width by a factor *N*. Both networks operate at equal voltages on the two transistors and this ensures the compensation of

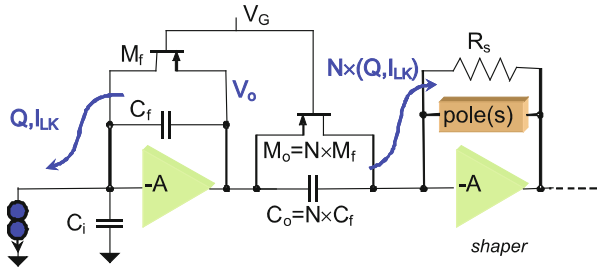


Fig. 10.30 Charge amplifier with continuous reset, pole-zero cancellation and compensation of non-linearity in the feedback transistor [10, 46]

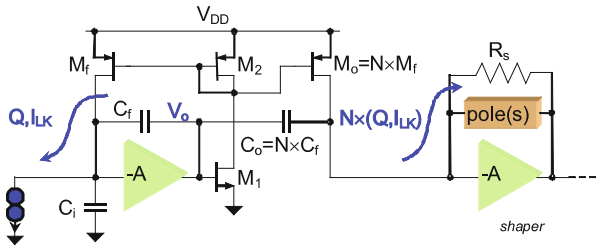


Fig. 10.31 An alternative configuration for a charge amplifier with pole-zero and transistor nonlinearity compensation [10]

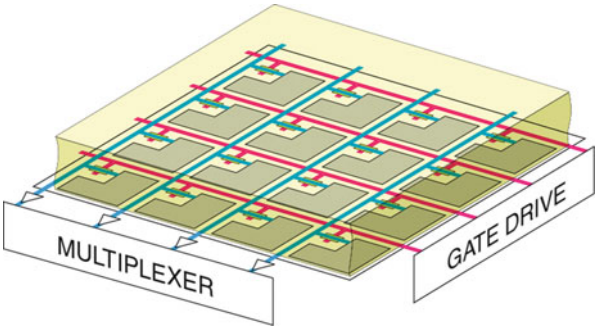


Fig. 10.32 Matrix readout of integrating pixel detectors. Transistors switches are integrated on the detector substrate

nonlinearity and pole-zero cancellation. The charge (current) gain in the first stage including the pole-zero network equals *N*.

An alternative configuration is shown in Fig. 10.31. An advantage of this configuration is that it separates the bias point of the transistors M_f and M_o from the virtual ground of the amplifiers. This results in a larger dynamic range. The configuration in Fig. 10.31 can be used as an input stage (charge amplifier) or as a second stage where several gain stages are needed. An analysis of both configurations is given in Ref. [10, 70, 71]. An overview of dc charge restoration circuits is given in Ref. [47].

10.8.3 Pixel Matrix Readout

Large pixel arrays can be conveniently read out by a matrix arrangement as illustrated in Fig. 10.32. The charge due to a photon or charged particle is stored on a pixel capacitance. The switches (one/pixel) connect a row of pixels to charge amplifiers located at the bottom of the columns. In this way, a multiplexing density is achieved between that for a separate readout for each pixel and the CCD with only

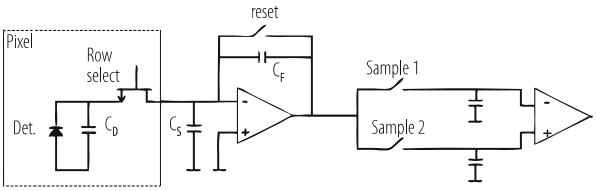


Fig. 10.33 Basic circuit diagram of the matrix readout with a switch (and no amplification) in each pixel [41–43]

one readout for the entire array. Such an array can be used in single event counting mode at sufficiently low event rates, or in charge integrating mode at very high event rates. This type of readout allows the use of the same technology, or different technologies, for the detector and the switching transistors. For many applications this approach is the best compromise between interconnect complexity and the speed of readout. An example of such an imaging detector for x-ray radiography is described in Ref. [41], and a silicon detector with Junction FET switches in Ref. [42]. The equivalent circuit diagram of such a readout with correlated double sampling is given in Fig. 10.33. It is convenient to group the noise contributions in this case in two classes: those that can be reduced by correlated double sampling and those that cannot. Referring to the notation in Fig. 10.33, the former are the reset (kTC) noise of capacitances C_f and C_s, and the latter are the reset noise of the charge collecting pixel C_d, shot noise from pixel dark current and amplifier series white noise. The reset (kTC) noise on the pixel capacitance C_d cannot be reduced by double correlated sampling according to the discussion in Sect. 10.7 and as illustrated in Fig. 10.27, because *both* the signal and the kTC noise start building up after the charge transfer from the pixel to the charge amplifier. Thus for the simplest matrix readout the minimum noise is limited by the pixel capacitance, Table 10.1, e.g., ~ 400 electrons rms for C_d ~ 1 pF at 300 K. There will also be a significant contribution by the amplifier noise if the connection capacitance along the column to the amplifier is large due to the size of the pixel array.

A much lower noise can be achieved in pixel arrays with "amplified pixels" as illustrated in Fig. 10.34. Each pixel has three or more transistors to perform the basic functions of reset, amplification and row selection. A CMOS matrix readout die can be bump-bonded (or in the future, directly bonded) to a pixel detector, or serve as a monolithic active pixel sensor (MAPS), Ref. [44]. Performing the reset in each pixel separately from the charge transfer (unlike in the simplest matrix readout without in-pixel amplification) allows almost complete cancellation of the pixel kTC noise by double correlated sampling, in applications where the time interval between the two samples is not too long (see Fig. 10.27). Noise levels in the range of 10–20 electrons rms have been achieved with small pixels (20–30 fF). The same level of noise has been achieved with larger pixels (~1 pF) and more conventional charge amplifiers and pulse shaping [48]. A lower noise with silicon pixel detectors has been achieved by integration of a field effect transistor on the high resistivity

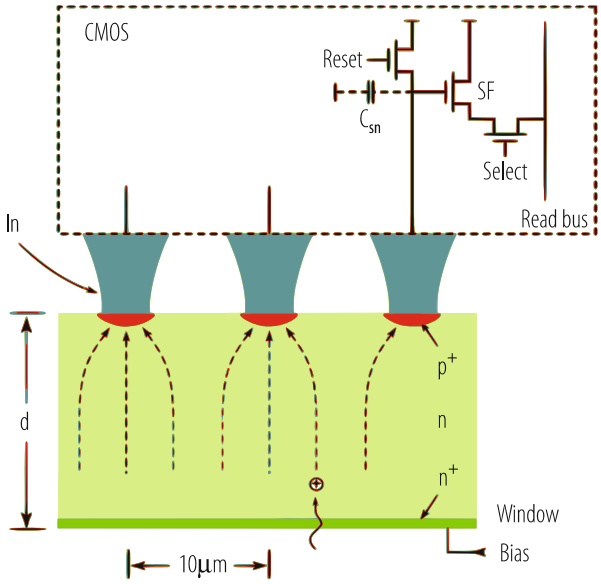


Fig. 10.34 Three transistor cell for readout of pixel detectors comprised of a source follower (SF), reset transistor switch and select transistor switch, illustrated for a bump-bonded hybrid detector. In monolithic active pixel CMOS detectors (MAPS), the readout cells are integrated with sensing diodes [44]. An overview is given in Ref. [43]

detector die (DEPFET, Ref. [49]). The lowest noise has been achieved with CCDs, Sect. 10.8.4. Each of these results was obtained after optimization for a particular application and each one involves a different set of parameters.

10.8.4 Charge Conserving (CCD) Readout

The charge coupled device (CCD), Fig. 10.35, is a device with the highest charge detection sensitivity among the photon and particle detectors. Low noise in the single electron rms range has been achieved [35, 69], thanks to a very low capacitance of the readout node and the integrated source follower (~15–30 fF), and the absence of any continuous conduction path to the sensing node. A switched reset is used after reading out the charge. Correlated double sampling is used to make the *kTC* noise negligible, as discussed in Sect. 10.7. Nearly complete *kTC* noise cancellation is achieved because the first sample is taken before the signal charge is transferred to the readout node (floating gate). Dual slope integration/trapezoidal weighting function is used for optimal filtering of the transistor (source follower) series noise, as discussed in Sect. 10.4.6.

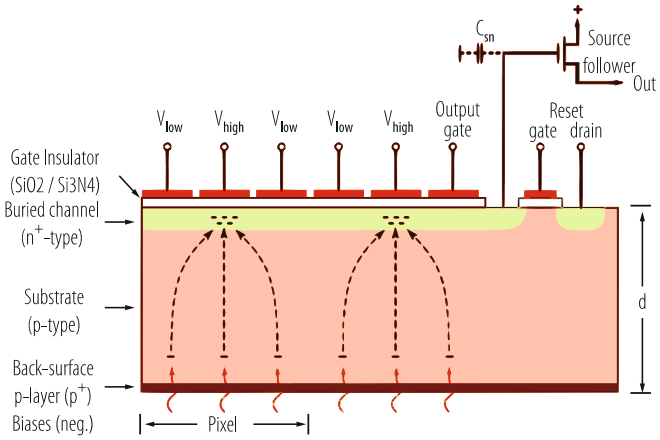


Fig. 10.35 An illustration of CCD read-out. The source follower (shown only schematically) is integrated on the CCD substrate as a buried channel enhancement mode MOSFET. The signal charge transferred from left to right is sensed at the floating electrode between the output gate and the reset gate. C_{sn} is the "node capacitance" = floating gate + gate to source capacitance of the transistor

10.9 Electronics Technology Outlook

10.9.1 Scaling of CMOS

The process of reduction of the size of CMOS transistors has continued for more than 40 years, and it is only very recently that it has started approaching fundamental physical limits. In this process the device is being scaled down in all three dimensions and in voltage, while the doping concentration is being increased. All dimensions are being reduced by a scaling factor α. This reduction includes all geometrical parameters of the device such as the gate oxide thickness t_{ox}, channel length L, channel width W, and junction depth. The substrate doping concentration is increased by the same scaling factor. The voltages applied were also expected to be reduced by the same scaling factor. These scaling rules were very clearly described in some detail already in 1974 [50], and the resulting device properties as a function of α are given in Table 10.2.

The scaling by a factor $\alpha = \sqrt{2}$ every ~2 years has followed the Moore's law [50] remarkably well until recently, with one exception. In the last few steps below the channel length $L \sim 0.13$ microns it is not possible to reduce the applied voltages by the same scaling factor. Downscaling has made the speed of the MOSFETs higher, the power dissipation per circuit lower, and it has enabled an ever-increasing level of integration (the number of transistors on a single chip). Aside from the enormous progress in all digital devices, it has made possible increasingly complex functions in the readout of detectors by integration of mixed signal (analogue and digital) circuits in Application Specific Integrated Circuits (ASICs).

Table 10.2 Scaling rules: dependence of device properties on scale factor α

Device property	Scaling rule
Electric field E	const.
Conductance (transconductance) $= I/V$	const.
Current density	const.
Power density	const.
Capacitance	$1/\alpha$
Speed $= I/CV = g_m/C = f_T$	α
Switching energy $= CV^2$	$1/\alpha^3$
Power/gate $= CV^2 f$	$1/\alpha^2$
Circuits density (transistors/unit area)	α^2

The principal impact on analogue ASICs has been:

- More, faster transistors
- Lower capacitance-lower noise readout of small detector pixels
- Better radiation resistance
- Prospects for vertical integration with high resistivity silicon detectors

However, the impact on analogue circuits has been very positive only up to a certain point of scaling. Scaling of CMOS into the deep submicron range (below 100 nm down to a few nm) has some undesirable consequences for low noise amplifier design:

- The low supply voltage "headroom" in scaled CMOS processes imposes limits on analogue circuit topologies. The increasing ratio of *VTH/VDD* rules out the use of many classical analogue design topologies.
- The *cascode connection*, useful in providing high gain loads and current sources, becomes difficult to realize once *VDD* falls below ~1.2 V.
- The electrostatic control of the channel by the gate is reduced, resulting in reduced ratio *Idon/Idoff*. In CMOS transmission gates, commonly found in sample/hold and switched capacitor circuits, self-discharge rates increase due to incomplete current cutoff.
- The reduced electrostatic control of the channel results also in a lower ratio of the drain conductance and the transconductance, i.e., the gain of the transistor. To achieve a certain gain more amplification stages may be needed.
- The dynamic range of capacitance based circuits, memories and sample & hold circuits, is reduced (the ratio of stored charge to the kTC noise), both due to the reduced V_{dd} and due to a lower storage capacitance.
- Gate tunneling current arising with thin oxides contributes shot noise.

While the scaling in the CMOS technology is driven by the entire information technology industry, some applications require higher operating voltages than dictated by the smallest feature sizes. This has been recognized, and the large semiconductor foundries offer options with a thicker gate oxide in their deep submicron platforms. For example, a nominal gate oxide thickness in the 65 nm platform (or "node") is about 65/50 $\approx$ 1.2 nm, but the process includes also on the

same wafer an oxide thickness of ~5 nm, that requires a minimum channel length of ~250 nm and allows a supply voltage of ~2.5 V. This will make possible in a single ASIC both high density digital circuits with the minimum feature size, and higher voltage input/output circuits and precision analogue circuits with an effectively larger feature size.

Analog design below 100 nm becomes gradually more difficult. The complexity of ASIC design rules and the costs of the design tools increase steeply as the feature size is reduced. This is discussed in some detail in Ref. [51].

While the developments in electronics technology have already made possible very large and complex detectors, the power dissipation associated with increasingly complex signal and data processing has been a problem and will remain a principal limitation and challenge.

Along with the quest in the CMOS scaling continuing down to the nodes in the ~7 to 28 nm range, an additional path to higher circuit densities and to higher speed of digital circuits (by shortening the interconnections) is three-dimensional integration of several thin layers of CMOS circuits [52, 53]. A significant breakthrough in particle and photon silicon detectors will be integration (by direct bonding) of a thick (50–500 μm) high resistivity p-i-n detector die to one or more thin (~10 μm) CMOS layers of readout electronics.

10.9.2 Transistors at Low Temperatures: Applications in Future Detectors

Most of the electronics for particle detectors has been based on silicon CMOS devices following the trends in electronics industry. Over the last two decades the technology of silicon–germanium heterojunction bipolar transistors (SiGe HBT) has been developed [54]. These devices have some key properties superior to the bipolar junction transistor (BJT), notably a much higher current gain and a much higher unity-gain frequency. The HBT, unlike the BJT can operate at liquid nitrogen temperature, and furthermore, with an increased current gain. Recently emerging cryogenic applications have generated renewed interest in low temperature properties of both CMOS and HBT technologies [54, 55]. As with HBTs, it has been found that various CMOS device properties improve at low temperatures. The mobility and the transconductance increase by a factor of two to three as the temperature decreases from 300 K to 43 K, and the (inverse) subthreshold slope decreases from ~90 mV/decade to ~20 mV/decade resulting in a higher ratio of the transconductance to the drain current. The transistor (series) thermal noise has been observed to decrease monotonically from 300 K to 40 K [56]. There is one *caveat*: CMOS transistors have to be operated (at any temperature) under conditions where hot electron generation, which is more pronounced at low temperatures, is minimized [57].

In particle physics, recently developed cold monolithic electronics has enabled scaling up of liquid argon time projection chambers to a very large size, required for studies of neutrino oscillations and nucleon decay. Such detectors, which originated with the ICARUS project [58], in the size range of up to few hundred tons, will have to be built in the range, unthinkable so far, of tens of kilotons. The number of sense electrodes (wires) and the readout amplifiers will be in the range of ~$5{\cdot}10^5$ to $3{\cdot}10^6$, assuming a reasonable electron drift length in the time projection chamber (TPC) and the sense wire length. Readout by monolithic amplifiers (with multiplexing) placed at wire electrode frames, results in a significantly lower noise (due to a lower capacitance) than with long cables bringing the signals from each sense wire to the outside of the cryostat, and in a much lower number of cables and feedthroughs. This also allows the design of the cryostat to be freed from the signal cable constraints. The "cold electronics" in this case will have a beneficial impact on both the engineering and the physics, i.e., a higher signal-to-noise ratio allowing better background rejection and a higher precision in the track measurement and reconstruction increasing the sensitivity for detection of interesting phenomena [59, 60]. Such a detector, on a smaller scale (~75 tons of liquid argon with ~8300 sense wires and electronic channels immersed in liquid argon), has been built and operated for two years. Uniformity and stability of the gain and noise has been demonstrated [61].

10.9.3 Beyond the Moore's Law

As the MOSFET technology advances further into the nanoscale domain (gate widths down to 5–10 nm), Moore's law, as defined for CMOS transistors, is running up against the physical, technical and economical limitations, and eventually against the granularity of matter (silicon lattice constant ~0.54 nm). Physical phenomena associated with small dimensions, such as quantum mechanical effects and fluctuations in the decreasing number of dopants, take place. These effects cause leakage currents (incomplete drain current turn-off and gate tunnelling currents) and dispersion in device parameters (threshold voltage). Associated with the necessity for tighter control of all device and fabrication process parameters are increasing costs.

Recent research toward smaller, faster and lower power devices has been concentrated on "beyond CMOS" devices [62–64]. As with the CMOS, new developments are being driven by the needs for high density-high speed digital and computer circuits. Besides the physical variables considered as "computational variables" familiar in CMOS (current, voltage, charge), other variables are being considered (electric dipole, magnetic dipole, orbital state). Among possible device concepts being explored are:

- Tunneling FET
- Graphene nanoribbon FET

- Bilayer pseudo spin FET
- SpinFET
- Spin transfer torque/domain wall
- Spin torque oscillator logic
- All spin logic device
- Spin wave device
- Nanomagnet logic
- III–V tunnel FETs

Each of these devices may have some properties superior to CMOS, but none, so far, satisfies the set of simple criteria that CMOS does. For example, upon analysis, the spintronic devices have longer switching delays and higher switching energies, due to inherent time of magnetization propagation.

Any "Beyond CMOS" device should have many of the same characteristics as CMOS devices:

- Power gain >1
- Ideal signal restoration and fan-out (output of one device can drive two or more devices)
- Feedback prevention (output does not affect input)
- High ON/OFF current ratio ~ $10^{5–7}$
- Low static power dissipation
- Compatibility with Si CMOS devices for mixed functions

The consensus about the future technology has been so far: No new device is yet on the horizon with a potential to completely replace CMOS. More likely, new devices may emerge by gradual evolution. New or special functions (e.g., memories, [34]) may become possible in the nanoscale devices by new physics and such devices may be merged into CMOS circuits to enhance overall performance. Impedance matching may be necessary from the quantum resistance values (kohms) down to the 50–100 ohm range. The overall logic operations and communications will still be based on CMOS. Future integrated circuits are likely to still contain a majority of CMOS devices with a few other beyond-CMOS devices performing various specialized functions. An in depth evaluation of the various device concepts under investigation is given in Ref. [65, 72].

Work on carbon nanotubes as active electronic devices and passive devices (interconnects) has been going on for the last three decades. Some interesting devices and phenomena have been described extensively in literature [64]. Carbon nanotube as a channel in an MOS transistor structure has higher mobility, reducing electron transit time. However, accurate placement of carbon tubes, and the need to form higher current channels by placing multiple carbon tubes in parallel, has been a challenge to fabricate uniform devices. Graphene based devices are analyzed in Ref [66].

In the quest for nanoscale devices and higher density of analogue and digital functions other limitations appear. Since all logic circuits (even spintronics circuits)

need electrical contacts at the terminals of gates and channels (source and drain) their size will be limited by the metallization dimensions.

As far as particle and photon detectors are concerned, further progress in digital circuit technology, based on nanoscale nodes, will result in increased functionality of the detector circuits, particularly those integrated with detectors, as long as they are economically available from multi-project foundry services or as commercial components. The development of multi-layer (3D) circuits for detectors has been challenged by limited access to this technology, given the relatively small quantities needed in physics experiments and high costs.

As far as the analogue front-end circuits for particle detectors are concerned, it appears that CMOS will be less useful below the ~65 nm node (Sect. 10.9.1 and Ref. [51]). In higher precision circuits, such as analogue to digital converters and switched capacitor memories, the dynamic range is limited by the low power supply voltages (one volt or less) at the upper end and by the *kTC* noise at the lower end. Fortunately. the provision of thicker oxide devices on the same nanoscale platforms leaves the choice of the operating voltage, the gate length and width to the designer (as discussed in Sect. 10.9.1).

An interesting domain, outside of the CMOS mainstream, has been presented by *single-electron transistors* (SETs), [67, 68]. Single electron transistors are devices with a capacitance so small that a single electron can generate a measurable voltage (above the thermal voltage). To be observable, the mean energy of the electron on the capacitance must be several times larger than its thermal energy (*kTC* noise). This sets an upper limit on the capacitance of the control electrode (gate) of the transistor (and the sensor connected to the transistor) and on the operating temperature. From Sect. 10.7 and Table 10.1, an upper limit for the capacitance at room temperature would be 1 aF (10^{-18} F). Alternatively, the temperature would have to be 300 mK, to allow a capacitance in the range of ~1 fF. SETs, although known for more than three decades, have not been considered suitable for integration due to large variability in their fabrication.

SETs might be suitable as very sensitive electrometers for *equally low capacitance* sensors. That sensitivity is for the most part due to their low capacitance and in part due to improved carrier transport properties. At 300 K they are not matched to even the smallest pixels of present particle and photon detectors, and will be useful only with an entirely new generation of very fine grained detectors operated at low temperatures.

Acknowledgments The author is indebted to his colleagues, Gianluigi De Geronimo, Paul O'Connor, Sergio Rescia, Bo Yu, and the late Pavel Rehak for many stimulating discussions and, over time, for contributions of material that has contributed greatly to this article. In particular, most of the induced signal simulations are due to Bo Yu, the circuits shown in Figs. 10.23, 10.30 and 10.31 originated in publications by Gianluigi De Geronimo and Paul O'Connor. Paul O'Connor provided much of the material on CMOS scaling, and Sergio Rescia worked with the author on *1/f* and *kTC* noise calculations. Special thanks are to Anand Kandasamy for help with editing of the manuscript and the figures.

References

1. G. Knoll, *Radiation Detection and Measurement*, John Wiley & Sons (2000).
2. C. Grupen and B. Shwartz, *Particle Detectors*, 2nd edition, Cambridge University Press (2009).
3. G. Lutz, *Semiconductor Radiation Detectors*, Springer-Verlag (1999).
4. H. Spieler, *Semiconductor Detector Systems*, Oxford University Press (2005).
5. I. Iniewski (ed.), *Medical Imaging*, John Wiley & Sons (2009).
6. D.M. Binkley, *Tradeoffs and Optimization in Analog CMOS Design*, John Wiley & Sons (2008).
7. C.C. Enz, E.A. Vittoz, *Charge-based MOS Transistor Modeling*, John Wiley & Sons (2006).
8. E. Gatti, P.F. Manfredi, *Processing the Signals from Solid-State Detectors in Elementary Particle Physics*, Rivista de Nuovo Cimento, 9(1) (1986).
9. V. Radeka, *Low-Noise Techniques in Detectors*, Annu. Rev. Nucl. Part. Sci. 38 (1988) 217.
10. G. De Geronimo, *Low-Noise Electronics for Radiation Sensors*. In: Medical Imaging, K. Iniewski (ed.), John Wiley & Sons (2009) p. 127.
11. W. Shockley, *Currents to Conductors Induced by a Moving Charge*, J. Appl. Phys. 9 (Oct. 1938) 635.
12. S. Ramo, *Currents Induced by Electron Motion*, Proc. IRE 27(1939) 584-585.
13. E. Gatti, G. Padovini, V. Radeka, *Signal Evaluation in Multielectrode Detectors by Means of a Time Dependent Weighting Vector*, Nucl. Instrum. Meth. 193 (1982) 651.
14. N.R. Campbell, *Proc. Cambridge Philos. Soc.* 15 (1909) 117.
15. P.N. Luke, *Unipolar Charge Sensing with Coplanar Electrodes – Application to Semiconductor Detectors*, IEEE Trans. Nucl. Sci. 42 (1995) 207.
16. V. Radeka, *1/|f|Noise in Physical Measurements*, IEEE Trans. Nucl. Sci. NS-16 (1969) 17.
17. C.M. Compagnoni et al., *Statistical Model for Random Telegraph Noise in Flash Memories*, IEEE Trans. Electron Devices 55(1) (2008) 388.
18. K. Kandiah, M.O. Deighton, F.B. Whiting, *A Physical Model for Random Telegraph Signal Currents in Semiconductor Devices*, J. Appl. Physics 86(2) (1989) 937.
19. P. van der Wel et al., *Modeling Random Telegraph Noise Under Switched Bias Conditions Using Cylostationary RTS Noise*, IEEE Trans. Electron Devices 50(5) (2003) 1378.
20. A. Konczakowska, J. Cicshosz, A. Szewczyk, *A New Method for RTS Noise of Semiconductor Devices Identification*, IEEE Trans. Instrum. Meas. 57(6) (2008) 1199.
21. L.K.J. Vandamme, F.N. Hooge, *What Do We Certainly Know About 1/f Noise in MOST?*, IEEE Trans. Electron Devices 55(11) (2008) 3070.
22. R. Wilson, *Noise in Ionization Chamber Pulse Amplifiers*, Philos. Mag., Ser. 7 Vol. xli, Jan. (1950) 66.
23. V. Radeka, *Optimum Signal Processing for Pulse Amplitude Spectrometry in the Presence of High-Rate Effects and Noise*, IEEE Trans. Nucl. Sci. NS-15 (1968) 455.
24. M. Konrad, *Detector Pulse Shaping for High Resolution Spectroscopy*, IEEE Trans. Nucl. Sci. NS-15 (1968) 268.
25. V. Radeka, *Trapezoidal Filtering of Signals from Large Germanium Detectors at High Rates*, Nucl. Instrum. Meth. 99 (1972) 535.
26. S. Rescia, V. Radeka, unpublished notes.
27. G.R. Hopkinson, D.H. Lumb, *Noise Reduction Techniques for CCD Image Sensors*, J. Phys. E: Sci. Instrum., 15 (1982) 1214.
28. E. Gatti et al., *Suboptimal Filtering of 1/f Noise in Detector Charge Measurements*, Nucl. Instrum. Meth. A 297 (1990) 467.
29. V. Radeka, *Semiconductor Position Sensitive Detectors*, Nucl. Instrum. Meth., 226 (1984) 209.
30. V. Radeka, *State of the Art of Low Noise Amplifiers for Semiconductor Radiation Detectors*, Proc. Int'l. Symposium on Nuclear Electronics, Versailles, 1968, Vol. 1 (1968) 46-1.
31. V. Radeka, *Field Effect Transistors for Charge Amplifiers*, IEEE Trans. Nucl. Sci. NS-20 (1973) 182; see also, V. Radeka, *The Field-Effect Transistor – Its Characteristics and Applications*, IEEE Trans. Nucl. Sci. NS-11 (1964) 358.

32. H.B. Callen, R.F. Greene, *On a Theorem of Irreversible Thermodynamics*, Phys. Rev. 86 (1952) 701.
33. H.B. Callen, T.A. Welton, *Irreversibility and Generalized Noise*, Phys. Rev. 83 (1951) 34.
34. D.B. Strukov, K.K. Likharev, *CMOL FPGA: A cell-Based, Reconfigurable Architecture for Hybrid Digital Circuits Using Two-Terminal Nanodevices*, Nanotechnology 16 (2005) 888.
35. R.P. Craft et al., *Soft X-ray Spectroscopy with Sub-Electron Readnoise Charge-Coupled Devices*, Nucl. Instrum. Meth., A 361 (1995) 372.
36. F.S. Goulding, D.A. Landis,, N.W. Madden, *Design Philosophy for High-Resolution Rate and Throughput Spectroscopy Systems*, IEEE Trans. Nucl. Sci. NS-30 (1983) 301.
37. V.T. Jordanov, G.F. Knoll, *Digital Synthesis of Pulse Shapes in Real Time for High Resolution Radiation Spectroscopy*, Nucl. Instrum. Meth. A 345 (1994) 337. X-ray Instrumentation Associates, *Appl. Note* 970323-1.
38. O. Benary et al., *Liquid Ionization Calorimetry with Time-Sampled Signals*, Nucl. Instrum. Meth. A 349 (1994) 367.
39. B.T. Turko, R.C. Smith, Conf. Record of the 1991 IEEE Nucl. Sci. Symp., (1991) 711.
40. O. Benary et al., *Precision Timing with Liquid Ionization Calorimeters*, Nucl. Instrum. Meth. A 332 (1993) 78.
41. N. Matsuura et al., *Digital Radiology Using Active Matrix Readout: Amplified Pixel Detector Array for Fluoroscopy*, Med. Phys. 26(5) (1999) 672.
42. W. Chen et al., *Active Pixel Sensors on High Resistivity Silicon and Their Readout*, IEEE Trans. Nucl. Sci. 49(3) (2002) 1006.
43. V. Radeka, *CCD and PIN-CMOS Developments for Large Optical Telescopes*, Proc. SNIC Symposium, Stanford, Ca., April 3–6, 2006.
44. G. Deptuch et al., *Monolithic Active Pixel Sensors with In-pixel Double Sampling Operation and Column-level Discrimination*, IEEE Trans. Nucl. Sci. 51(5) (2004) 2313.
45. V. Radeka, *Signal, Noise, and Resolution in Position-Sensitive Detectors*, IEEE Trans. Nucl. Sci. NS-21(1) (1974) 51.
46. G. De Geronimo, P. O'Connor, *A CMOS Fully Compensated Continuous Reset System* IEEE Trans. Nucl. Sci. 47 (2000) 1458.
47. G. De Geronimo et al., *Front-End Electronics for Imaging Detectors*, Nucl. Instrum. Meth. A 471 (2001) 192.
48. P. O'Connor et al., *Ultra Low Noise CMOS Preamplifier-Shaper For X-ray Spectroscopy*, Nucl. Instrum. Meth. A 409 (1998) 315.
49. J. Kemmer, G. Lutz, *New Semiconductor Detector Concepts*, Nucl. Instrum. Meth. A 253 (1987) 365.
50. R.H. Denard et al., *Design of Ion Implanted MOSFETs with Very Small Physical Dimensions*, J. Solid-State Circuits SC-9 (1974) 256.
51. L.L. Lewyn et al., *Analog Circuit Design in Nanoscale CMOS Technologies*, Proc. IEEE 97(10) (2009) 11687.
52. IBM J. Res. & Dev., Issue on *3D Chip technology*, 52(6) (2008).
53. Proc. IEEE, Issue on *3-D Integration*, 97(March) (2009).
54. J.D. Cressler, *On the Potential of SiGe HBTs for Extreme Environment Electronics*, Proc. IEEE 93(9) (2005) 1559.
55. T. Chen et al., *CMOS Reliability Issues for Emerging Cryogenic Lunar Electronics Applications*, Solid-State Electron. 50 (2006) 959.
56. G. De Geronimo et al., *Front-End ASIC for a Liquid Argon TPC*, IEEE Trans. Nucl. Sci., Vol. 58, No. 3, June (2011) 1376–1385.
57. S. Li et al., *LAr TPC Electronics CMOS Lifetime at 300 K and 77 K and Reliability Under Thermal Cycling*, IEEE Trans. Nucl. Sci. Vol. 60, No. 6, Dec. (2013) 4737–4743.
58. S. Amerio et al., *Design, Construction and Tests of the ICARUS T600 Detector*, Nucl. Instrum. Meth. A 527(2004) 329, and references therein.
59. V. Radeka et al.,*Cold Electronics for "Giant" Liquid Argon Time Projection Chambers*, Journal of Physics: Conference Series, Vol. 308, No. 1, (2011) 012021.
60. H. Chen et al.,*Cryogenic Readout Electronics R&D for MicroBooNE and Beyond*, Nucl. Instrum. & Meth. A 623 (2010) 391–393.

61. R. Acciarri et al., *Noise Characterization and Filtering in the MicroBooNE Liquid Argon TPC, JINST,* 12:P08003, 2017, 1705.07341.
62. T. Hiramoto et al., *Emerging Nanoscale Silicon devices Taking Advantage of Nanostructure Physics*, IBM J. Res. & Dev. 50(4/5) (2006) 411.
63. M. Haselman, S. Hauck, *The Future of Integrated circuits: A Survey of Nanoelectronics*, Proc. IEEE 98(1) (2010) 11.
64. Int. J. High Speed Electron. Syst., Issue on *Nanotubes and Nanowires*, 16(4) (2006).
65. D.E. Nikonov, JSNM 21, 497 (2008).
66. F. Schwierz, *Graphene Transistors: Status, Prospects, and Problems,* Proc. IEEE, 101(7) (2013) 1567.
67. D.D. Smith, *Single Electron Devices*, Int. J. High Speed Electron. Syst. 9(1) (1998) 165.
68. M.H. Devoret, R.J. Schoelkopf, *Amplifying Quantum Signals with the Single-Electron Transistor*, Nature 406(31 Aug) (2000) 1039.
69. R.P. Kraft et al., *Soft X-ray Spectroscopy with Sub-Electron Readnoise Charge-Coupled Devices*, Nucl. Instrum. Meth. A 361 (1995) 372.
70. G. De Geronimo, P. O'Connor, *A CMOS Detector Leakage Current Self-Adaptable Continuous Reset System: Theoretical Analysis*, Nucl. Instrum. Meth. A 421 (1999) 322.
71. P. O'Connor, G. De Geronimo, *Prospects for Charge Sensitive Amplifiers in Scaled CMOS*, Nucl. Instrum. Meth. A 480 (2002) 713.
72. D. E. Nikonov and I. A. Young, *Overview of Beyond-CMOS Devices and a Uniform Methodology for Their Benchmarking,* Proc. IEEE, 101(12) 2013.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 11
Detector Simulation

J. Apostolakis

This chapter provides an overview of particle and radiation transport simulation, as it is used in the simulation of detectors in High Energy and Nuclear Physics (HENP) experiments and, briefly, in other application areas. The past decade has seen significant growth in the availability of large networked computing power and particle transport tools with increasing precision available to HENP experiments, and enabled the use of detailed simulation at an unprecedented scale.

After describing the uses of detector simulation and giving an overview of its components, we will examine selected cases and key uses of detector simulation in experiments and its impact.

11.1 Overview of Detector Simulation

11.1.1 Uses of Detector Simulation

Simulating the generation of particles in an initial collision, the interaction of these primaries and their daughter particles with the material of a detector and the response of the detector is a key element of recent experiments. Its importance has grown with each generation of experiments from LEP, B-factories, and the Tevatron, through to the current generation at the LHC, due to the increased precision requirements. It will be a significant element in planned experiments at super-B factories, the International Linear Collider (ILC), the Future Circular Collider (FCC), and also for numerous ongoing and smaller future experiments.

J. Apostolakis (✉)
CERN, Geneva, Switzerland
e-mail: john.apostolakis@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_11

Simulation serves many purposes at each point in the lifecycle of an experiment or a facility. Different types of simulation are used, typically with an increasing level of detail during a lifecycle. At first, fast simulation follows only the most energetic particles, typically the particles arising from the primary interaction, through simplified geometrical descriptions to obtain average values for energy deposition in the key volumes in which a signal is generated. Later, in more detailed simulation, the interactions create secondary particles, which in turn are tracked and create further descendant particles; interactions are treated with more detail and energy deposition includes fluctuations. This enables the estimation of many quantities including the measurement resolution of detectors, and correlations.

To prepare a proposal for an experiment, different versions of the setup are simulated. For each design a simplified version is constructed and simulated, typically using a tool such as SLIC [1] or DDG4 [2] which provide templates for detector components and hit generation. Putting this together with tailored digitization and reconstruction, many key characteristics of a design necessary for technical design report [3] can be evaluated.

The energy dispersion or resolution of calorimeters, the longitudinal and lateral leakage, corrections to momentum measurements, the backscattering of particles into trackers or other 'upstream' components can be estimated for different designs. Accurate simulation is a powerful quantitative tool for optimizing the performance, the size and cost of each sub-detector. In addition, its use extends to quantifying the tradeoffs between the performance of combinations of detectors, and finally for optimizing the global performance of a complex modern detector.

During the prototyping and calibration phase of a detector, simulation is utilized for test beam setups of single detectors or combinations to ensure that their performance is understood and can be accurately modeled. The accuracy expected in today's high precision experiments requires agreement between simulation and test beam measurements at the level of 1% or better. This type of high-quality quantitative comparison is the basis for evaluating through simulation significant corrections to measurements in the experiment that cannot be obtained in a test beam or directly from in-situ data once the experiment starts operating.

The possibility for detailed modeling of the conversion of energy deposition into signal within a detector is a key strength of detector simulation. This requires detailed knowledge of many detector-specific effects. Simulation is an important tool also for the data analysis phase of an experiment. After its accuracy has been validated against single-particle benchmarks and test beams, an experiment's simulation can be utilized to model tracks of different types, and even full signal and background events. Detailed quantitative aspects of simulated tracks are used in preparing methods for particle identification and measurement before the start of an experiment, and continue to be crucial in many methods that utilize data to calibrate complex quantities such as the jet energy scale. It can also be used in modeling the separation of signal from background contributions to measured data.

The impact of a performant detector simulation, whose predictions matched data, was demonstrated in the first years of the operation of the LHC experiments, during

which reconstruction and calibration in many channels were achieved within the first years of operation, a fraction of the time of previous hadron collider experiments [4].

A challenging aspect of the use of simulation in data analysis is the estimation of the systematic errors of the simulation. The accuracy of a simulation depends on the accuracy of the description of the detector's geometry and material composition, the knowledge of the beam or primary particles delivered and the intrinsic accuracy of the simulation tool(s) used. The detector simulation tools, in turn, are limited by several factors: the availability and known accuracy of measurements utilized to tune or validate the physics models, in particular of the cross-sections; the limitations of the physical models in reproducing the energy spectra and other properties of interactions; the approximations utilized to obtain adequate computational speed, to simulate the required number of events using the available computing resources.

There is an important tradeoff between the level of detail, both in the geometrical description of a setup and the choice of physics modeling options, and the computational cost of large-scale simulation. In the past 5 years the LHC experiments have been able to use detailed simulation to produce several billion events per year [5] providing unprecedented support of analysis in hadron collider physics. The increase in luminosity in the HL-LHC era will bring a need for much higher statistics of simulated events, whereas projected growth in computing power is forecast to be modest in comparison [6]. This is driving research to achieve substantial performance gains in full simulation in GeantV [7], and the expectation that faster approximate simulation will be relied upon once again for most analyses—leading to efforts to create new kinds of fast simulation which more accurately capture additional features of the full simulation, including fluctuations of key quantities.

11.2 Stages and Types of Simulation

- Event generators and detector simulation
- Scale from full detail to fast simulation
- Simulation of energy deposition or signal generation
- Assessment of radiation effects
- Key tools: Event generators, detector Monte Carlo, radiation transport
  - Detector Monte Carlo: GEANT, FLUKA, GEANT4
  - Radiation related MC: FLUKA, MARS, MCNP/MCNPX
  - Signal generation: Garfield

The simulation of the passage of particles through a detector and the response of the detector's sensitive elements typically proceeds in different stages. In the first stage an external event generator simulates the initial interaction and then decays short-lived particles; the results are the "primary" tracks. The second stage is detector simulation and involves the tracking of the primary particles through the structures of the detector, sampling interactions with their components, and creating

secondary particles. In the third stage the "hit" information is processed, to estimate the signals that result.

In detector simulation the secondary particles and their descendant particles are tracked in turn. Information about the passage of particles through the sensitive parts of the detector is recorded as 'hit' information. For tracking detectors usually the individual position, momentum and particle type or charge information of each track is recorded; in calorimeters, the energy deposition within a cell is kept as a sum over all tracks. A key characteristic of detector simulation is that tracks are treated independently. Each particle created is tracked in turn, until all have exited the setup, suffered a destructive interaction, or have been dropped as unimportant according to a user's chosen criteria. The dropping can be triggered, for example, when the energy of a track falls below a threshold or arrives in a particular 'unimportant' region. Potential indirect interactions between particles are not treated as part of the detailed simulation. As such, the creation of space charge in a gaseous detector must be introduced in an experiment's 'user' code or else treated separately.

11.2.1 Tools for Event Generation and Detector Simulation

The creation of the primary particles by the high-energy interaction is modeled using specialized event generators. The type of interaction, energy range and applicability of these generators differ significantly: whether they include hadron–nucleus and/or nucleus–nucleus interactions, or the type of physics beyond the standard model they provide. Typically event generators are independent programs: including the established PYTHIA [8] and FRITIOF [9], which use the Lund fragmentation model [10], and more recent ones such as HERWIG [11]/HERWIG++ [12]. Most provide users with tunable parameters and the ability to create sets of parameter values ('tunes') compatible with the most relevant reference data at the energies of interest.

Some Monte Carlo tools include high-energy event generators: e.g. DPMJET is available in FLUKA, and has been used to simulate ion–ion collisions at RHIC and the collisions of high-energy cosmic rays in the atmosphere.

Codes for the simulation of the detector must handle geometries of significant complexity and a large number of volumes and they must model the full set of hadronic, electromagnetic and weak interactions as accurately as required, potentially within constraints of CPU time. In the past 20 years different tools have been used for this purpose, including GEANT 3, FLUKA and GEANT4. Other multi-particle codes for particular applications including the MARS [13, 14] code, and the SHIELD code which focus on ion–ion interactions. Different codes share some physics models; for example PHITS and MARS share several models with MCNPX, an extension of the neutron-gamma gold-standard code MCNP.

GEANT version 3 [15] was utilized by LEP experiments (ALEPH, L3), the TeVatron experiments at Fermilab, numerous other experiments and also by the ALICE experiment at LHC as its main simulation engine. It includes detailed descriptions of electromagnetic interactions down to 10 keV. For hadronic physics it

relies on external packages: GHEISHA [16], GCALOR [17], which uses CALOR89 [18], and GFLUKA, which interfaces to the 1993 version of FLUKA [19].

FLUKA [20], after a major overhaul in the 1990s, offers microscopic models for 60 elementary particles, all types of ions at energies from 1 keV to 10,000 TeV/A. It has been used for detector simulation by the Opera and ALICE experiments, in radiation assessment, accelerator collimation and target tuning, and many applications beyond HEP. A key emphasis and strength have been its single, consistent, core hadronic model, PEANUT, with a Dual Parton Model (DPM) based high energy cascade above 5 GeV, a generalized intranuclear cascade and suite of models for the excited nucleus. For nucleus–nucleus interactions above 5 GeV, it utilizes interfaces to DPMJET-III [21] event generator for interactions. There is an option for the detailed treatment of neutrons down to thermal energies, which uses the multigroup approach involving energy bins, and weighted averages of cross sections and interaction production. Physics processes for electromagnetic interactions and lepto-nuclear interactions are included. FLUKA focuses on a single set of physics processes, which are curated and validated extensively by its authors. A small set of optional variations of physics processes are provided, e.g. for the simulation of low-energy neutrons. The majority of uses in HEP lie outside detector simulation. Examples include the estimation of radiation backgrounds in experimental areas, whether in accelerator facilities or underground halls, and the modeling of beam interactions with collimators in accelerators. Extensive studies of the LHC radiation environment have been carried out using it over the past decade, and a FLUKA model of the full LHC collider is the production simulation for radiation studies.

GEANT4 [13] is the basis of the simulations of BaBar, ATLAS, CMS, LHCb and a large number of smaller experiments. Its standard configuration provides electromagnetic interactions for charged particles and gammas down to 1 keV, hadronic processes for nucleons, mesons and ions, models of electro-nuclear, lepto-nuclear interactions, radioactive decay of nuclei and optical processes for photons at visible and near-visible energies. A variety of hadronic processes has been used to span all projectiles at energies up to 1 TeV, with recent extension to 100 TeV for Future Circular Collider (FCC) applications [22]. An option for neutron interactions from 20 MeV down to thermal energies is available using cross-sections for individual elements and isotopes (a technique called 'point-wise'). GEANT4 takes a toolkit approach, enabling and requiring its users to choose the parts required for their application area, including the configuration of physics models. Recommendations of physics model configurations are provided for several established type of application and for a number of HEP and external application domains; validations for several HEP applications have been undertaken in collaboration with experiments. For other application domains users are invited to undertake the appropriate validation, potentially using their specific data, and interacting with GEANT4 experts.

A few other codes provide extensive modeling of multiple particle types, including the PHITS [23] code and MCNP family. The most recent MCNP version 6 [24, 25] was created from the merger of MCNP5, which focused more on traditional neutron-gamma applications including simulation of nuclear facilities

and reactors and the all-particle offshoot MCNPX. Its models will be contrasted with the capabilities of FLUKA and GEANT4, but it has not seen use in HEP detectors, due to lack of electron/gamma models above 1 GeV, restrictive licensing and export control.

In order to compare with measurements, the response of detectors to the energy deposited by an event's tracks must be estimated. One tool used for the detailed estimation of the energy deposition in a gas detector is Garfield [26]. It generates low-energy gammas and electrons, down to eV energy using HEED [27], uses the electric field calculated externally, and transports electrons and ions under the combined influence of its electromagnetic field and diffusion. The recent Garfield++ rewrite [28] and extension extended its capabilities, added refined electron transport and physics models for semiconductors, and enabled interfacing with GEANT4 [29]. Its computational cost is 2–3 orders of magnitude larger, so it is utilized sparingly: for studies and to generate an accurate parameterization of a detector's response [30] for use in large scale simulations.

11.2.2 Level of Simulation and Computation Time

The modeling of every physical interaction, from the initial particle's energy down to the interaction of eV scale photons and electrons—or even the interactions of neutrons down to thermal energies—is possible. The computing cost of such simulation is prohibitive for most practical applications, and simplifications are required. Yet in some cases it is necessary to simulate down to very low energies, for example in order to estimate the activation of materials by neutrons.

In complex detectors, such as in an LHC detector, the full simulation of each event takes between 0.1 and a few minutes on modern computers, depending on the type of event (minimum bias or t/t-bar) and the region simulated (rapidity coverage). This limits the number of events that can be simulated.

In some applications the simulation effort can be reduced for many events: by simulating first the particles that are involved in the trigger. Otherwise, one may seek to limit the number of secondary particles generated or the total track length simulated, or to simplify the treatment of the most frequent interactions.

Another alternative is fast simulation. This involves selecting only a fraction of tracks for simulation, and approximating the detector and key physics interactions in order to reduce the computation time per event by one, two or more orders of magnitude. Fast simulation is a powerful tool for modern experiments, as it allows speedy turnaround to address changing conditions or assumptions, and to explore different model parameters at an affordable computing cost. It can be calibrated using full simulation, data or both. However it is not capable to estimate resolutions and correlations, and it can be harder to obtain systematic errors.

Recently ATLAS has created a hybrid simulation mode by selecting for detailed simulation the conical regions of the detector around the most energetic primary particles, and using fast simulation models for the remainder [31].

11.2.3 Radiation Effects and Background Studies

The background in a modern detector can be due to many factors, including remnants of past events, accelerator generated backgrounds, and the backscattering of particles by the detector's surroundings. These can require simulation, in order to determine their level and characteristics. Also in many cases effects of the experiment on its surroundings or its constituents such as activation must be estimated. Simulation is an essential component.

Tools that are utilized for these tasks include FLUKA and the MARS code [32]. In addition to inclusive physics models, where the whole interaction is simulated, MARS contains exclusive models, where the leading particles and a sample of other secondaries is produced by an interaction.

Biasing is a technique in which some 'unlikely' trajectories are enhanced by a large numerical factor and assigned a weight inverse to this factor, in order to rapidly estimate their effect. It is an essential component of background applications. In many cases a result cannot be obtained without it; in others it improves greatly the accuracy of the result. Good statistical accuracy can be achieved within a fraction of the computation time required for an unbiased, so-called 'analogue', calculation for means and similar observables. Correlations, widths and other second order observables can be obtained only in some cases and by recording key additional information during simulation.

MARS has been used for accelerator and background calculations for many facilities [14] and experiments [33]. FLUKA also has seen wide application in this domain.

11.3 Components of Detector Simulation

- Geometry description and navigation
- External fields
- Electromagnetic physics models
- Hadronic physics models
- Low-energy neutron interactions
- Accuracy of simulation
- Fast simulation
- Signal generation
- Biasing, production thresholds

A complete tool for simulating particle interactions and detector response must include the description of a detector's geometry and material, the input or selection of primary particles, the modeling of all relevant physical interactions and the extraction of information such as the energy deposition and particle passage (hits).

Most tools also account for the effects of external electromagnetic fields on charged particles, provide visualisation of the geometry and simulated events. They provide for tallies, output of key physical quantities calculated during the simulation, such as totals of energy deposition, dose, particle flux and fluence. They also provide the opportunity for the user's code to filter and record track quantities at each step.

The geometry module provides the ability to describe the material composition and the geometry of the setup in terms of volumes. The tool must be able to navigate inside this volume description, identify the volume in which a point is located and calculate the distance to the next boundary in a given direction. The capabilities of the geometry modeler determine the type of volumes and their relative placement: whether volumes are generated directly as finite shapes, or whether they are the result of the intersection of surfaces; whether all volumes must be placed within a single 'world' volume, or a hierarchy of volumes can be created. In order to simulate a large, complex detector with hundreds of thousands to millions of volumes, the geometry modeler needs to support hierarchical geometry definitions.

To ensure good performance the key geometry operations must be computationally efficient; in particular, the computation of the distance to intersecting a boundary is critical. Optimisation methods which rely on data precomputed at initialization inspired by ray-tracing are used to greatly reduce computation time wherever many candidate sub-volumes exist.

Some experiments have chosen to use a geometry modeler external to the simulation tool. They use the same geometry description and modeler inside a Virtual Monte Carlo framework [34]. This interfaces to different simulation tools for modeling interactions: GEANT, FLUKA and GEANT4; they are labeled 'physics engines' and can be selected at runtime.

11.3.1 External Fields

The effects of external electromagnetic fields on the trajectory and energy of a charged particle track are modeled utilizing the Lorentz equation. The equations of motion for the position, the momentum and optionally the polarisation of the particle are integrated to obtain the position and state of a particle after a distance *s*. In special cases, such as a constant magnetic field, an analytical solution can be used. In the general case, numerical integration is used, typically with a Runge-Kutta method.

After integration the idealized curved path of a particle track in a magnetic field is propagated through the geometry of a detector. The curved trajectory is split into linear chord segments, which are used to navigate in the model geometry. The intersection of a chord segment is progressively refined to identify the location where the curved path crosses a geometry boundary.

11.3.2 Introduction to the Transport Monte Carlo Method

At each step of a simulation, the Monte Carlo method for particle transport needs:

- the cross sections in the current material of each physical interaction;
- an algorithm to select which interaction occurs next;
- a method to apply the effects of each interaction: to generate new particles, and change the state of the potential surviving projectile.

The Monte Carlo method [35], general techniques [36] and its application to particle transport for charged and neutral particle transport [37] are well described in the literature. We touch on a few of the essential features.

A key ingredient is a source of 'pseudorandom' numbers, distributed uniformly in an interval, usually [0,1). These are obtained from pseudorandom number generators [38] and must come with guarantees of non-correlation, such as those provided by the generators based on ergodic theory, MIXMAX [39, 40] and RANLUX [41, 42], or at least have survived a barrage of empirical tests [43] to suggest there are no correlations which affect the Monte Carlo estimates.

For a general particle the total interaction cross-section (summed over all interactions)

$$\sigma_{\text{total}} = \Sigma_i \sigma_i$$

is used to sample the step length s, using a random number r from the interval (0,1):

$$s = -(1/\mu)\ \ln\ r$$

where the absorption coefficient μ is proportional to the cross-section σ_{total} and density ρ. Thereafter, the type of interaction that will occur at this step is chosen. The probability for one particular type of interaction to occur (in one step) is proportional to its cross section.

In the ideal case, all interactions would be sampled this way. However, in practice a different approach is needed, as the cross section diverges for the emission of soft photons and delta rays. A systematic treatment proposed by Berger [44], separates collisions that alter the state of the particle below a chosen threshold, typically for the momentum transfer. These are not sampled individually; only their collective effect is sampled. The collisions above the threshold are simulated individually.

In this approach, the part of the cross section corresponding to an interaction below this threshold is labeled the continuous part, and it does not contribute to limiting the step. Its effect is applied separately as an integral over the length of the step, to the state of a track.

The discrete part of an interaction contributes its cross-section to limiting the step

$$\sigma = \sigma_{\text{discrete}} + \sigma_{\text{continuous}}$$

Its cross section represents all interactions resulting in secondary particles with energy E above the threshold energy E_0:

$$\sigma_{discrete} = \int_{E_0}^{\infty} \frac{d\sigma}{dE} dE$$

This treatment is required for the Bremsstrahlung process and delta-ray production, due to the large number of secondaries produced with low energy.

11.3.3 Electromagnetic Interactions and Their Modeling

The modeling of physical processes can be separated into models of electromagnetic (EM) interactions, models of hadronic interactions (involving the strong nuclear force) and the decay of unstable particles mediated by the weak force. The Monte Carlo simulation of EM interactions of charged particles with atoms has been well established in HEP applications since the advent of EGS4 in the 1980s. EGS4 was able to produce and track photons, electrons and positrons down to 10 keV.

At typical HEP energies of 1–100 GeV, the number of particles of the electromagnetic shower is large. The full simulation of all resulting particles is costly in computational resources, and a selection of particles is undertaken to represent the shower. Typically, particles are tracked until they reach a certain energy threshold, the tracking cut, and discarded. In addition, secondary particles are emitted only if their energy is above a chosen energy, called the production threshold. For specific applications the high density of energy deposition near the endpoint of a track (Bragg peak) is relevant, and can be simulated.

Electromagnetic interactions of gamma-ray photons include Compton scattering, the photoelectric effect and, gamma conversion, the production of electron–positron pairs. Cross sections for each process are calculated directly from theoretical or empirical formulae, or parameterized. For example the Klein-Nishina formula is used for the cross section of Compton scattering. To improve execution speed the value of the cross sections are pre-calculated at several energies; the value at any other energy is obtained by interpolation.

The method used to model multiple scattering, in particular near boundaries, is a key feature of a simulation tool. Obtaining accurate results using less computing power, and obtaining results that are stable when varying parameters (such as the production threshold or tracking cut) are significant algorithmic challenges.

The EGS approach for the simulation of photons and electrons and its implementation were pioneered by Nagel. It was improved and shared within the HEP community as the EGS3 [45] and EGS4 [46] code systems. From these other HEP codes for EM interactions are descended, or inspired.

The underlying assumptions in Monte Carlo simulation of radiation transport are the same amongst these and modern codes: materials are assumed to be amorphous,

and beam particles do not interact. The methods for modelling transport of photons and electrons used in Monte Carlo codes are based on sampling of differential cross sections obtained from approximate theoretical calculations. A recent review offered a comprehensive description of the principles and approximations [47] for models of electron and photons up to 1 GeV, documenting widely used models and those of the precise modern electron–photon code Penelope [48].

11.3.4 Interactions of Photons

Photon interactions are 'discrete' interactions, that occur at a point and can be modeled this way. This makes them much simpler than modeling the interactions of charged particles. Interactions considered including photoelectric, Compton incoherent scattering, electron–positron pair production and potentially Rayleigh coherent scattering.

The cross section for each interaction is sampled from measured or theoretical distributions. In some cases a simplified form is used, to reduce the cost of computation with a simplified description of the energy and Z dependence. Else, the values for each material at particular values can be pre-computed and stored in tables for interpolation.

Once the type of interaction is chosen, its products are sampled from the appropriate distributions. Pseudorandom numbers are used to sample the energy, angles and momenta from the differential distributions [37]. The original particle's state, if it survives, is altered to preserve energy and momentum.

The interaction of energetic photons with nuclei is discussed in the hadronic section below. Often specialized tools are used to simulate optical photons and their collection. It is possible, though, to generate optical photons and model reflection, refraction and absorption on different types of surfaces. GEANT4 is able to do this.

11.3.5 Interactions of Charged Particles

The simulation of the electromagnetic (EM) interactions of charged particles is complicated by the large cross section for elastic interactions and of ionization, which produces low-energy electrons (delta electrons).

In a few cases it is useful to simulate every single interaction of a charged particle in a medium, including its elastic collisions with nuclei, the ionization of atoms and creation of delta electrons, and the 'hard' interactions, which create photons or electron–positron pairs.

All production simulation tools estimate the cumulative effect of the elastic scattering off nuclei. It is modeled in several different ways. Many utilize the multiple-scattering approach pioneered by Goudsmit and Sanderson [49] as their basis. One simple way to sample angular deviation over a short step is Moliere's

theory [50], which is used in GEANT 3, but is limited to small angles. The approach of GEANT4 borrows from Lewis' description [51].

Key effects of multiple scattering are angular deflections and straggling. The latter's most important effect is the shortening of the distance travelled in the direction of the initial momentum. This must be modeled in order to obtain correct energy loss for the passage through material, as is done in GEANT 3. The second effect of straggling, the displacement in the lateral directions is correlated with angular deflection. Similarly to EGS4, FLUKA and GEANT4 also sample this displacement, each using different algorithms. This enables longer steps while maintaining accuracy. The best algorithms allow longer steps, or more accurate modeling of the correlations between the affected changes in the state of the particle.

The algorithm for multiple scattering has a significant effect on the results obtained in many detectors and setups. Examples include the partition of energy in sampling calorimeters, the correlation between the deflection of muons and their positions after substantial material. In particular, many quantities are very sensitive to the details of its formulation and implementation. These include the fraction of low-energy electrons ($T << 1$ MeV) scattering backwards at the interface between low and high Z materials and the correlation between the direction of a particle exiting a detector (e.g. muon) and its position.

In addition to the sampling of the final state, high accuracy for electron transport necessitates careful treatment of multiple scattering of low-energy electrons at boundary crossing [52]. New algorithms have been developed for exact electron transport without special treatment for boundary crossing [53, 54]. These algorithms have been implemented in electron–photon Monte Carlo codes: PENELOPE [55], EGSnrc [56] and EGS5 [57]. A comparison [58] in 2007 benchmarked the algorithms in several of these codes and in GEANT4, using data from custom setups with thin slabs. PENELOPE and EGSnrc demonstrated the best performance, while GEANT4 obtained good results only with specific settings.

The new GEANT4 GS model [59], available since GEANT4 release 10.3, implements Kawrakow's approach to provide angular deflections for any size step without free parameters and offers the option for accurate boundary crossing. It achieves the best agreement, amongst GEANT4 models, with a wide range of benchmark data including backscattering data.

Models for specialized processes, such as transition radiation, exist in some tools including GEANT4 [60].

11.3.6 Hadronic Interactions and Their Modeling

In contrast to the simulation of EM physics processes, the simulation of hadronic physics processes from first principles is not possible, except partially at the high energy limit. At all energies, the cross-sections and the models used are based, directly or indirectly, on measured data of hadron–nucleon and hadron–nucleus interactions, and on phenomenology.

The most common particles produced by hadronic interactions are nucleons, pions and kaons. The diversity of particles and interactions make modeling a great challenge. Specialised codes including HETC [61], GHEISHA, CALOR [18], FLUKA, and SHIELD [62] were developed for HEP and other application areas in the 1970s and 1980s. Few models and codes from other application domains have been available which cover a substantial part of the energy range (above the 1–2 GeV used in spallation applications) and the full set of particles needed for HEP. One exception is the MCNPX tool, an extension of the MCNP code for neutron/gamma radiation transport and reactor simulation. MCNPX hadronic models are shared with MARS.

Interactions are divided into elastic and inelastic, which produce new particles in the final state. The smaller cross-sections for inelastic interactions of hadrons with nucleons compared to EM interactions, and the growing multiplicity and variety of particles emitted in interactions above a few GeV, result in significantly different structure for hadronic showers.

The modeling approach depends on several factors: the availability of detailed experimental measurements; the complexity of final states of reactions for a particular combination of incident particle, energy and target; and the availability and suitability of theoretical or phenomenological descriptions. In many application domains there is a requirement for conservation of energy, momentum and quantum number in each interaction, and for the coincidence or correlation between the products of an interaction; in selected cases conservation of energy only as an average over different interactions may suffice. A small number of interaction models, including GHEISHA, and most low-energy neutron interactions sampling methods treat particle interactions only in the average, and do not conserve energy and momentum.

For many applications full energy conservation of individual interactions and the treatment of the correlations of particle tracks is required in order to obtain reliable results. For example, the estimation of the energy resolution of hadronic calorimeters is strongly affected by these factors.

In many cases a phenomenological model is supplemented by fits of model parameters with available data. In a few cases (evaluated) data libraries are used directly—typically for low-energy neutron transport. Another approach is to use parameterizations, either directly of data or indirectly for the parameters of simplified models, as in GHEISHA.

11.3.7 Models of Interactions at Low Energies

At the lowest energies, the largest hadronic cross section belongs to the elastic interaction, which is a coherent interaction of a hadron projectile with the full nucleus. Hadronic cross sections, including those for elastic scattering, are typically parameterized from data.

Inelastic interactions, which excite the nucleus, typically become relevant at energies of order MeV. They are modeled with a statistical approach for energies up to about 100 MeV. The original Weisskopf evaporation model [63] describes the emission of protons and neutrons from nuclei in thermodynamic equilibrium. It is supplemented by several additional de-excitation channels, which compete to occur. These include the multi-fragmentation model for highly excited nuclei [64], Fermi breakup of light nuclei, fission of heavy nuclei and photon evaporation.

In an alternative approach, following the Generalised Evaporation Model (GEM) of Furihata [65], nuclei with up to 28 nucleons are evaporated directly. This improves greatly the description of the emission of light and medium fragments, with an extra computational cost. GEM is an option in MCPNX, in GEANT4 and a similar approach is used in FLUKA. Fragments heavier than ^{4}He, though emitted infrequently, are important for specific applications, such as the response of silicon devices and damage to them; using the GEM approach is recommended for these.

At energies up to about 300 MeV a simple algorithm can be used to count the number of excited nucleons and holes of missing nucleons in the Cascade Exciton Model (CEM) model [66]. Such models are called pre-compound (or pre-equilibrium) models. A pre-compound model of this type is implemented in GEANT4. And CEM.03 [67], which is included in MARS, MCNPX and MCNP 6, is an improved CEM descendant.

These models are also used to calculate the de-excitation of nuclei after interactions at the higher energies, important in many applications. For example, they determine the energy that is lost to nuclear breakup, and the partitioning of energy between the low-energy protons, neutrons and gammas that are produced. These processes produce the majority of neutrons, whatever the initial interaction, and as a result affect substantially the escaping energy, lateral shower profiles and compensation of calorimeters—amongst other observables.

11.3.8 Cascade Models of Hadron–Nucleus Interactions at Intermediate Energy

At energies above about 100 MeV an intranuclear cascade model is used for nucleon and pion projectiles. In a cascade the interaction is modeled as a succession of independent collisions of the projectile (and secondaries) with individual nucleons inside the target nucleus [68].

In a cascade, the nucleus is described in two ways. It can be an ensemble of nucleons positioned at random locations, sampled from a model of nuclear density—as used Quantum Molecular Dynamics (QMD) models such as UrQMD [69], in the GEANT4 Binary cascade model [70] and in the Liege cascade INCL. Else, the nucleus can be composed of a number of shells of constant density, as in the original cascade of Bertini, in FLUKA, and in INUCL and its descendant, the

GEANT4 Bertini-type cascade [71]. A correction factor is used for the depletion of nuclear shells by earlier interactions in both FLUKA and GEANT4 Bertini.

It uses 'free-space' cross-sections derived from hadron–proton measurements, or, in some cases cross-sections modified for the presence of the nuclear medium [72]. FLUKA accounts for nuclear medium effect on the Δ resonance properties in the treatment of pion interactions [20].

Hadrons may move in curved trajectories according to a chosen nuclear potential, as in GEANT4 Binary and FLUKA, or in straight lines (GEANT4 Bertini). In both cases, the potential is used to update the momentum of all hadrons before interactions. Interaction products can use models or be sampled from data, must observe the Pauli exclusion principle, and are subject to a hard-core nucleon repulsion. Particles arriving at the nuclear boundary with enough energy are ejected, while others are reflected and continue to interact. The difference of the total energy of the remaining nucleons and the ground state energy that corresponds to them is the excitation energy.

After either a fixed time or once the excitation energy has dipped below a threshold, the remaining nucleons are handed to a pre-equilibrium or de-excitation module. A pre-equilibrium model, such as the Precompound model, can eject higher energy nucleons and is used with the GEANT4 Binary cascade. A similar model is used in FLUKA. The subsequent de-excitation module combines evaporation, Fermi-breakup for light nuclei, fission for heavy nuclei and other competitive channels.

A common de-excitation module is shared by all models in FLUKA. A custom simpler de-excitation module is used in GEANT4 Bertini; recent extensions enabled it to use the default "Preco" Pre Compound and de-excitation module, used after Binary cascade and the higher energy string models (QGS and FTF). In MCNPX, MARS and MCNP a common module is used with different parameters by the CEM cascade and the higher energy LAQGSM [67] model.

Most cascade models are expected to work up to 1.5–3 GeV, yet they can provide good results from 30 MeV up to 3–5 GeV. At higher energies their assumptions break down, because quark degrees of freedom become important. Including additional reactions with larger multiplicities of products, and effects such as formation time (a simplified treatment of quantum-mechanical effects as a time interval before secondary hadrons can interact) allows a cascade model to have a higher energy limit. This is the case for the GINC/PEANUT cascade in FLUKA and the GEANT4 Bertini cascade.

Early versions of GEANT4 until 9.6 also included the CHIPS model [73] applicable for intermediate energies. It described a nucleus in terms of nucleon clusters, and interactions as exchanges of quarks and was part of physics lists used in LHC Run 1.

The GEANT4 Bertini cascade underwent a substantial rewrite and upgrade [74]. As a result, all long-lived hadrons can be projectiles (adding K, Λ, Σ, Ξ and Ω). It also implements gamma- and lepto-nuclear reactions. Its energy range was extended up to 15–20 GeV, with the addition of final states with higher multiplicities, up to nine for proton–proton. Total and partial cross sections and final states were obtained from the CERN-HERA data compilations, and completed using symmetries and

general principles for unmeasured reactions and energies. The number of nuclear shells varies from one for the light, three for medium, and a maximum six for the heaviest elements.

As its main use is the simulation of LHC experiment detectors, a number of modeling and implementation choices were made to optimize computation speed. These include linear interpolation for the sampling of partial cross sections and large 10-degree bins in angular distributions—justified by the smoothing effect of additional interactions.

The newest cascade model in GEANT4, the INCL Liege cascade, is the one under the most active development. This event generator, under development since the 1990s [75], was developed to reproduce spallation data for reactions at 100 MeV–1.5 GeV using a parameter-free model. It interfaced with the ABLA code for deexcitation. Its original Fortran version up to version 4.6 [76] has interfaced to MCNPX and MCNP6. Recent development focused on the re-engineered INCL++ [77], which reproduced the performance of 4.6 and was extended to handle light ion projectiles (up to carbon16) and to higher energies, up to 15 GeV producing multiple pions.

Large suites of benchmark data are used to tune and verify the modeling of each cascade, covering neutron, proton and pion production at energies from 60 MeV to 3 GeV. Spallation data from inverse kinematic reactions on hydrogen targets at GSI with a range of projectiles from ^{56}Fe [78] to ^{238}U [79] at 1 GeV/nucleon provide different challenges for modeling and complement these data.

Comparisons of hadronic and in particular cascade models have been undertaken periodically under the auspices of the IAEA [80]. These benchmarks use a large set of thin target data to probe the accuracy and predictive capabilities of each model. INCL was found to be one of the competitive models. Similar test suites are utilized as a part of internal benchmarks and for tuning of models.

The details of the intranuclear cascade and the pre-equilibrium model determine all the emission of higher energy particles, but the details of the coupling and the quality of the de-excitation module are also critical to the performance of many applications from activation to calorimeter simulation. Good modelling of the resulting nuclei is required to ensure that the energy lost to nuclear breakup, a key component of non-compensation in hadronic calorimeters, is accurate. Recent results from RD52 [81] were interpreted as deficiencies in the modelling of nuclear breakup in GEANT4 version 9.4, showing that there is room for further improvement in this energy range.

11.3.9 High-Energy 'String' Models

Models for interactions at high energies (above 5–10 GeV) simulate quark-level interactions and rely on phenomenological descriptions of soft QCD interactions to generate low-energy hadrons from the remnants of the high-energy collisions. They are applicable to all hadron projectiles. Three variants are available: the Dual Parton

model [82] is implemented in DPMJET [21] and used in FLUKA. The Quark Gluon String (QGS) Model [83] in different variants is used in MARS/MCNPX, GEANT4 and in the QGSJET event generator [84]. A third model, the Fritiof model, is used and extended in the FTF model in GEANT4.

Hadrons are produced in the initial collisions and the decay of QCD strings, tubes composed of compressed gluonic fields [10] generated by the separation of colored quarks. Models implement a Lund-like string. The proto-hadrons generated by string decay, once past their formation time, together with outgoing nucleons, and the remaining cluster of nucleons are handed to a cascade for additional scattering in the nucleus (e.g. in FLUKA and in GEANT4 QGS_BIC and FTF_BIC physics lists using the Binary cascade), or passed directly to a pre-compound module (in the GEANT4 QGSP_BERT and FTFP_BERT physics list).

Developments over the last 10–15 years include a high-energy extension of the PEANUT cascade in FLUKA to undertake reactions handled previously by DPMJET; the connection of the Binary cascade in GEANT4 with the QGS model (as QGS_BIC) to re-scatter the slow products of high-energy model. More recent developments include the extensive improvement of the Fritiof-based FTF model in GEANT4 to model light anti-nucleus interactions at low energies and at rest [85], and to include an internal Reggeon cascade, and changes in the production ratio of different types of diquarks [86].

11.3.10 Treatment of Low-Energy Neutron Interactions

Amongst particles created by hadronic interactions, neutrons survive a longer time and are among those which travel the furthest. Also, they are amongst the most numerous. This makes their treatment important for many applications and correspondingly expensive computationally. Most neutrons are emitted in the de-excitation phase of a reaction and have energies of order MeV.

Neutrons produced in high-energy interactions (above 20 MeV) also lose energy in elastic and inelastic interactions (which release protons, alpha particles or light ions—or create gamma rays) before being captured by nuclei. Only part of this process occurs on fast timescales and others are much slower (μs–ms). By tracking time, it is possible to emulate the time dependence of the signal. This also allows the simulation tool to use a time threshold, and abandon neutron tracks after this time, in order to save computation time.

Neutrons' contribution to the visible energy measured in a detector comes via the transfer of energy to charged particles and from capture and other reactions with nuclei that generate gammas. Elastic scattering is particularly important in organic scintillators, where interactions with hydrogen transfer significant parts of energy and momentum to the recoil proton.

In some cases, are treatment of neutron interactions at a greater level of detail is required, potentially down to thermal energies. In HEP it is needed in special cases, such as the study of activation of materials for radiation safety purposes.

The detailed treatment of energy deposition in scintillators may also require a more detailed treatment of neutrons than provided by the simpler interaction models.

Simulating neutrons below 20 MeV relies on measurements of cross-sections for key processes, which have been assembled into established data libraries. Libraries, such as JEFF [87], ENDF/B VII [88], ENDF/B VIII [89], JENDL [90] and CENDL [91], include evaluations of cross sections and distributions of secondaries for key reactions based on a combination of measurements and estimates from nuclear model codes. They cover all measured interactions, from inelastic and capture at low energies; through inelastic interactions resulting in the emission of one or two neutrons (plus gammas); to multi-neutron production at tens of MeV. For some elements, data is available for several individual isotopes, while for others only the values for the natural composition are measured. Cross sections have many resonances in the keV–MeV region, complicating precise treatment.

For reactions that result in more than two outgoing particles, in most cases only spectra are available; i.e. information on the correlations of products is not included. As a result, sampling secondaries for one interaction in a way that conserves energy and momentum requires complex algorithms and additional computation. This has recently been introduced in PHITS; other codes rely on uncorrelated sampling and conserve energy only on average.

This detailed treatment consumes significant CPU and memory, because the full set of cross sections for all isotopes of all elements is required. Variants of this approach are utilised in MCNP/MCNPX (the gold standard for neutron simulation) and in the GEANT4's NeutronHP package.

A simpler approach averages cross sections over chosen sets of nuclei and fixed bins of energy. This 'multi-group' approach provides savings in memory use compared with the detailed approach and is adopted in FLUKA, as the option for precise treatment of low-energy neutrons. Accuracy is determined by the number of sets of nuclei and of the intervals of energy—and also the choice of the grouping.

In addition to the purely electromagnetic interactions of charged particles, and the interactions of hadrons with matter, it is necessary also to simulate the interaction of electrons, positrons and gammas with nuclei which result in hadronic final states. These photo-nuclear and electro-nuclear interactions account for a small portion of the total cross-section of gammas or electrons, below one percent at its peak. Yet they are the only interactions that convert electromagnetic energy into hadronic final states in typical HEP experiments.

Models for photo-nuclear and electro-nuclear interactions are provided in all multi-particle codes discussed [92].

Given the diversity of hadronic interactions, there is a need to focus on essential aspects. What most influences the accuracy of the description of the energy response, energy resolution of calorimeters and missing energy for hadronic calorimeters?

For HEP applications (in particular calorimeters) some of the key features are:

- most energy is deposited by low-energy particles, and its spectrum is independent of the type of projectile particle [93];

- the production of π^0 particles and the fluctuation of the energy fraction in this channel which leads to prompt EM energy deposition plays a determining role in the resolution of calorimeters [94, 95]; these π^0's can be the result of charge exchange or other hadron–nucleon collisions, or formed by the soft fragmentation, e.g. in the QCD string view of high energy reactions;
- the simulation of neutron generation, transport and interaction, which contribute to prompt and delayed signal, activation, and escaping energy;
- a component of missing energy (that influences the resolution) is the energy lost to nuclear breakup. The accuracy of the modeling of all hadronic reactions, but in particular the de-excitation stage, determines the quality of its simulation;
- the simulation of leading particles which determines the shower profile—the profile of energy deposition—and in particular the sharing between longitudinal compartments and the amount exiting in the direction of the projectile.

The accuracy of the modeling of the time dependence of different interactions, both in a tool and in an experiment simulation, are also essential.

The fluctuation in the fraction of energy going into each type of secondary particle (gammas, charged hadrons, neutrons, neutrinos) in all reactions is an essential feature of a simulation tool. It is important for predicting and modeling the resolution and other aspects of detector performance.

11.3.11 From Full to Parameterized ('Fast') Simulation

In order to obtain high statistics, it is necessary for some applications to utilize simulation which is much faster (typically by two orders of magnitude) than the detailed full simulation. There is a spectrum of such simulations with different approximations and compromises.

The coarseness of simulation, from detailed to fast, is determined by a number of variables: whether the geometry is described approximately or in great detail; whether secondary particles are generated from the interaction of primary particles; the degree to which particles are eliminated during tracking—for example, the relation between the energy of primary particles and an energy threshold; and the type of physics models utilized. Choosing the least level of detail and following only the primary tracks distinguishes the fastest simulation. In other variants, some aspects are simulated in more detail in order to obtain more precision. Different ways can be used: simulating more particles, adding physics or geometry volumes.

In the first two LHC runs, it has been possible to produce billions of simulated events using detailed simulation. In some experiments the forward sub-detectors were simulated faster, e.g. use by ATLAS of frozen showers for forward calorimeters. Projections for the High Luminosity LHC era foresee an order of magnitude gap between the statistics possible using GEANT4-based detailed simulation on projected 2025 hardware and the needs of most analyses. This gap is driving the continued development of fast simulation methods, and the research into methods

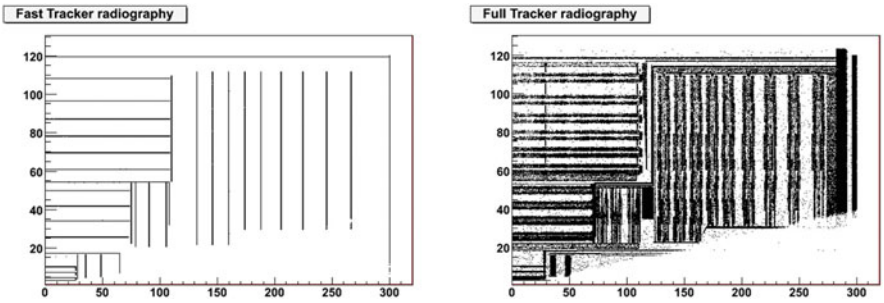


Fig. 11.1 Simulated radiography of a quarter of the CMS tracker geometry using the fast simulation FastSim (left) and full GEANT4-based simulation (right). This demonstrates the simplification of the geometry description used in fast tracker simulation, which projects the material onto smaller cylindrical shells, yet reproduces the hit structures and reconstructed tracks with the accuracy required in physics analyses. (Reproduced from [97])

for speeding up particle transport simulation. The GeantV [7] R&D aims to produce a prototype to demonstrate whether the core simulation work can be redesigned for more efficient use of current and forecast computer architectures with complex CPUs and deep memory architectures, with the goal of a speedup factor between 2 and 5.

The fastest type, so-called 'parametric' simulation involves the simplified propagation of tracks coupled with a reconstruction of fixed efficiency and idealized EM and hadronic calorimeters are given an input resolution. In one modern incarnation, Delphes [96], it is coupled with built-in reconstruction, and can be used to obtain first level estimates for some physics analyses from a simplified model of a detector. This type of simulation is used in the first feasibility stages of detector design, and to obtain a first understanding of physics analyses.

A more accurate type of fast simulation uses a simplified geometrical setup of a tracker device and/or sampling of showering using a parameterized distribution of a calorimeter to generate energy deposition hits and reconstruct events using about 100 times less computing resources than the full detailed simulation. An example of the simplified tracker geometry can be seen in Fig. 11.1, where the geometry of the CMS fast and full simulation of the tracker are compared.

The detailed geometry and physics of the full simulation, e.g. using GEANT4, is typically used as a yardstick for comparison of relevant physics quantities required by physics analyses, and sometimes to generate a library of pre-simulated showers at set energies for use in recreating realistic showers.

The LHC experiments have developed many different types of fast simulation, both specific to one part of a detector (tracker or calorimeter), and spanning the full detector.

One example is the mixing of detailed simulation for parts of the detector with simplified treatment within fixed regions or regions that depend on the particles inside an event or collision.

Recently ATLAS has produced an integrated simulation framework ISF [98], which includes detailed simulation and per-subdetector fast simulations. This enables the use of common modules for all elements of the simulation. A hybrid simulation mode is part of its design that selects conical regions around the most energetic primary particles for detailed simulation, and uses fast simulation models for the remaining parts of the detector. This capability has not yet been used in physics analyses.

11.4 Machine Learning for Fast Simulation

A completely new approach to parameterized simulation has emerged recently, exploring the potential of machine learning. One research avenue attempts to generate patterns of energy deposition that reproduced the distributions including fluctuations of key physical observables.

This has been the topic of interest and recent investigations using the generative-adversarial network (GAN) approach to the fast generation of patterns of hits. A first demonstration in a three layer LAr sampling calorimeter [99] was tested over an energy range from 1 to 100 GeV and a single particle direction. A potential speedup of O(100)–O(10,000) was demonstrated. Most physical observables of interest for the classification and calibration of tracks were well reproduced, but a few showed clear differences. This promising avenue will clearly be an area of significant research in the next years.

11.4.1 Accuracy of Simulation

The accuracy of the simulation is determined not only by the artificial differences or defects introduced by such simplifications, but also by intrinsic factors. These factors include the accuracy of the cross-sections for particular interactions and the capability of the physical models. These can be explored by comparing with experimental data.

Criticisms of detector simulation focus on key limitations and question the predictive power of hadronic interaction modeling for use in designing and tuning hadronic calorimeters [100].

11.4.2 Signal Generation

In order to model the signal produced in a gaseous detector, all processes that contribute to the generation of charge and its collection in the cathode must be simulated. The detailed simulation of a small number of events modeled in full detail is used to understand the characteristics of a detector. For large-scale simulation, a

simple model or parameterization is produced for the signal generation given the energy deposition.

The simulation involves a level of detail beyond other Monte Carlo simulation for HEP detectors. Some of the important aspects include:

- modeling the generation of all secondaries, without an energy threshold, in every single inelastic collision in the gaseous volume;
- the effects of elastic collisions in the transport;
- for efficiency, pre-calculating the convolution of the effects of the resulting diffusion and the drift in the electromagnetic fields of the detector;
- the effect of potential build-up of charge on detector elements, e.g. the space charge in a gas.

Due to the need to simulate down to the eV scale, this simulation requires detailed knowledge of the excitations of the molecular constituents of the mixture. Specialised programs are necessary for this simulation. Garfield calculates the electric field in many regular cell geometries. Then it combines it with the energy deposition for each atomic shell via a specialized Photo Absorption Ionization model integrated from Heed [101] to generate all secondary gammas and electrons. Transport of the charged particles in the electromagnetic field is coupled with diffusion, using pre-calculated transport coefficients generated by the Magboltz code [102].

It is possible to simulate an avalanche near a sensor wire in order to accurately model the signal arriving at the detector's electronics. This is typically required only for the detailed understanding of the effects of the shape of the signal and the integration characteristics of the electronics, and requires three orders of magnitude more computation than the simulation of the energy deposition in the gas volume. Alternatively, a fraction of the track can be simulated, in order to determine key characteristics, such as the arrival time of the signal. This reduces the computing requirements by about an order of magnitude.

It is also possible to calculate the effect of charged particles on an integrated circuit element [103]. Using the energy deposition to create electron–hole pairs, an external Technical Computer Aided Design (TCAD) program simulates the detailed response of the circuit. Applications of this technique have focused on the simulation of single event upsets [104], in which a cosmic ray track results in the flipping of a bit in a silicon circuit. Typically, the circuit response involves proprietary TCAD programs.

11.4.3 Production Thresholds and Other Biasing Techniques

For many setups, computing resources for simulation in full detail are not available. For example, large-scale experiments can require millions of events in order to establish patterns related to rare processes, yet the full simulation of events is prohibitive. In this case the choice must be made how to discard particular tracks in

order to achieve the required computing time per event, while influencing the most important results as little as possible.

The simulation time for each particle type is proportional to the number of steps, and typically the energy of a particle (so long as it does not escape the setup). The large number of electron and gamma particles in an EM or hadronic shower necessitates that these tracks are key to reducing the computation time. For this, either the treatment of each track must be simplified, or the average track length must be greatly reduced, or the number of particles tracked must be reduced. The use of a tracking cut reduces the average track length, and the use of a production threshold reduces the number of tracks simulated.

A number of methods are widely practiced to reduce the computing time by simulating only the more important particles: to generate only particles whose energy is above a threshold energy (production threshold); to kill tracks once they fall below an energy threshold (tracking cut), or to treat neutrons via their average cross section (multi-group).

In other setups the interest is to estimate the fraction of particles passing through or around a shielding barrier, which e.g. could stop all but one particle in a million. In such cases, a method to speed up the simulation is needed. To estimate the flux of particles passing through such a barrier the transport mechanics must be changed. Most changes will favor paths which have already crossed part of the barrier or are likelier to cross—e.g. because they have higher energy [105]. There are many methods to achieve this, all of which assign a weight to a particle track. Most involve the creation of extra copies of tracks or the killing of tracks. Some of the most common are importance biasing; leading particle biasing and weight window.

Importance biasing involves separating the geometry into regions of high and low numerical importance. At the boundary between such regions particles that go from low importance to high importance are enhanced in number (splitting), and their weight is reduced in proportion. Particles moving in the opposite sense are reduced in number (Russian roulette) and each one's weight is increased.

Leading particle biasing involves sampling the results of an interaction, favoring particles that have the highest energy (and most chance to penetrate) while sampling other particles in a representative way. In all cases surviving particles from populations, which are suppressed, are given higher weight in proportion to the difficulty of survival. On the other hand, enhanced particles (where two particle tracks are created from a single one) are given reduced weight. A single event can be split into a large number of 'histories', trial tracks that carry a different weight. Physical observables must be estimated accounting for the weight of a track—which can be interpreted as a probability:

$$\langle O \rangle = \frac{1}{N} \sum w_i O_i$$

The mean value of an observable O when using event biasing is calculated using the weighted sum of the values for each particle track 'history' i, which contributes, and the total number N of events (or trial histories).

11.5 Case Studies

The discovery of the top quark [106, 107] involved detector simulation only in a minimal and indirect way. Detector simulation was undertaken by CDF and D0 in the optimization phase of the design of their detectors. In D0 it informed the design of the interface between the central and end calorimeters [108]. Subsequently, simplifications were made for the detector simulation used in production, to reduce the computing time per event. The individual calorimeter plates were replaced by a large block, which contained a mixture of the absorber and active material. Optionally the response of particles below 200 MeV was parameterized. Comparisons with test beam determined that the full detailed plate setup agreed well; the simplified mixture setup was found to agree less well with test beam measurements, but judged adequate for most purposes.

Events simulated using the D0 simulation program were used in the later observation of the production of single top quarks [109]. However, the detector simulation was found to have significant limitations; these appear to have stemmed both from the simplifications of the modeling of the detector and its response and from the intrinsic limitations of the simulation tool. Several corrections were required, including a factor for the efficiencies of the trigger reconstruction, and for the efficiency to identify and select particles and jets.

In the LHC experiments, simulation was first utilized to model the response of the calorimeters to muon, electron and hadron beams. One key application has been the detailed calibration of the energy response of the electromagnetic calorimeter, to obtain an estimate of the energy of the incoming particle as a function of the inferred visible energy measured as signal in the different calorimeter compartments.

The ATLAS calorimeter system is complex, utilizing different detector materials, geometrical structure and technologies for the different rapidity regions. The insensitive material between parts of a detector distorts the energy signal due to tracks crossing this region. Simulation is used to obtain correction factors for the energy of the incident particle or jet.

ATLAS has undertaken an extensive comparison of test beam results with Monte Carlo simulation. Simulation utilising GEANT 3.21 in the 1990s was used for detector design studies and the first test beam comparisons. Progressively from 2000 onwards, test beam results were compared mostly with GEANT4, and in 2004 a Geant4-based simulation was declared the official ATLAS 'production' detector simulation in 2004.

Comparisons were undertaken first with the test beam results for individual detectors. A typical comparison started with muons, as minimum ionizing particles, to make a first verification of the material description and geometry of active parts. Next, measurements of electrons were compared, verifying the detector description of passive parts. This determines the factor for conversion of the deposited energy into the signal measured by each detector, its electromagnetic scale, by comparison with a beam of a particular energy, typically 100 GeV. Key

observables compared include the linearity of the response, the energy resolution and the shower longitudinal and lateral shapes.

The comparisons identified areas where improvement was necessary: there were problems in the large dependence of the electron energy response on the production thresholds. These were also reported in studies related to the use of GEANT4 in medical applications [110] and were corrected in GEANT4 release 8.0.

11.5.1 Calibration of EM Calorimeter Using Monte Carlo

The calibration of the ATLAS liquid argon (LAr) Electromagnetic Calorimeter for electrons was developed utilizing a detailed Monte Carlo simulation of the detector and its test beam [111]. The simulation included a very detailed description of the complex geometry, collection of the energy deposition into hits and conversion to digitized signal.

The simulation was shown to describe well all the relevant measurements, including the mean reconstructed energy, the distributions of energy deposition in particular compartments and the energy profiles in the longitudinal compartments and lateral sections. For example, the mean reconstructed energy was described to within 2% in the Pre-Sampler and the first two (of three) compartments of the accordion. Also the distribution of the total reconstructed energy is described very well—within the uncertainty due to the upstream material ahead of the setup and in front of the PreSampler.

Based on this agreement, the simulation was used to correct for several effects, which could not be measured or could only be estimated indirectly. One involved the average energy deposited in the dead material between the pre-sampler and the first compartment; this provided an estimate of this energy deposit. Another effect was Bremsstrahlung in upstream material, due to which a fraction of events arrives at the detector with reduced energy. A quantitative description was made and compared with the measured total energy deposit.

Dedicated Monte Carlo simulations were undertaken to study the systematic uncertainties induced by each effect. As a result, the reconstructed energy response in the energy range from 15 to 180 GeV was found to be linear within 0.1% (an exception is at 10 GeV, where it was found to be 0.7% lower). The systematic uncertainties due to incomplete knowledge of the detector, the test beam, or the reconstruction were found to be about 0.1% at low energies and negligible at high energies. The effect of the non-linearity at about 40 GeV and above on the measurement of the W^+/W^- mass was found to match the aimed precision of 15 MeV—provided it can be extended from the section tested to the full calorimeter.

11.5.2 Hadronic Calorimeters: Comparisons with LHC Test Beam Results

For the hadronic calorimeters, after muons and electrons, the test beam results with pions were compared to simulation. This served to validate the simulation of hadronic interactions, since the conversion of the energy deposition into signal is common with the electron test beams. Additional observables compared included the longitudinal and lateral shower shape.

The ATLAS Hadronic Barrel Calorimeter's (TileCal) use of scintillator requires the accurate simulation of neutrons for those interactions that contribute significantly to energy deposition and which occur within its time window of 150 ns. This is treated in the simulation.

ATLAS has also undertaken extensive test beam measurements of the response of the TileCal to pions. The most recent comparisons in the Combined Test Beam setup (2004) involved pion energies up to 350 GeV. These tests [112] examined the energy response and resolution of the calorimeter, and compared them with the predictions of a simulation based on GEANT4 (version 9.1).

Events were selected based on several criteria, including the energy deposition in a cryostat scintillator (SC1) placed before the TileCal. This cut was made in order to enable comparisons with the previous test beam. Potential biases from this selection cut on the response and resolution were studied. An approximation of this cut was used in the simulation: the energy deposition in the surrounding dead material. The change in the energy response due to this cut ranged from -2.5% to $+0.5\%$ depending on the energy and eta value (0.25–0.65) of the pion beam. This was reproduced within 0.5% at low energy (20 GeV) and within 1.5% at high energy (300 GeV).

The energy resolution was affected in a range from $+10\%$ to -10% between low and high energies respectively and reproduced within 2% for a large combination of angles and energies, except for a 4% deviation at one angle at 20 GeV. Comparing the final results for the energy response, the agreement obtained is within 3% for the full energy range studied (20–350 GeV). Typically, agreement at the 1–2% level is achieved for beam energies of 50–250 GeV; greater deviations are seen at 20 and 300 GeV (the latter, in particular, remains to be understood). For the energy resolution agreement at the 10% level is obtained.

The measurements of the reconstructed energy in the Atlas TileCal at energies from 20 GeV to 180 GeV have also been compared with simulation [113]. Figure 11.2 shows the ratio of simulation and data for the energy for the case of incident pions using different configurations of physics models in 2010 with GEANT4 version 9.2. Agreement for the mean energy between simulation and data ranges from -10% (the legacy LHEP physics modeling) to $+3\%$ (the production physics list QGSP_BERT). The root mean square deviation (RMS) of the reconstructed energy agrees within -4% (QGSP_BERT) to about $+15\%$ (LHEP).

Since a 1 MeV neutron could travel only 3 cm within 100 ns, even if it never interacted, the propagation of low energy neutrons—and thus their contribution to

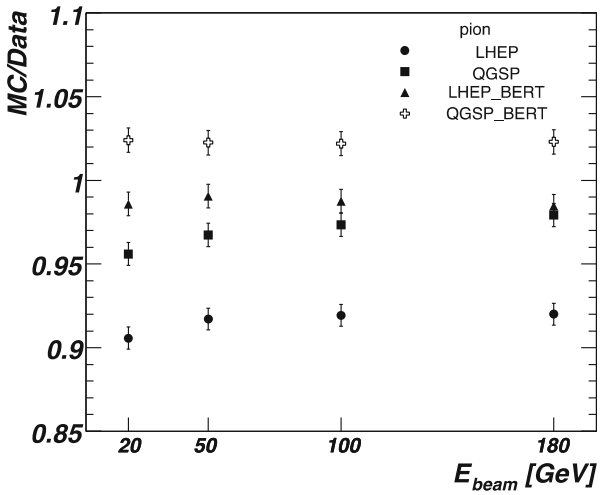


Fig. 11.2 Comparison of the reconstructed energy in the ATLAS Tile calorimeter (TileCal) with several different physics models. The ratio of the simulated (MC) response with the response reconstructed from test beam runs (Data) for pions of energies from 20 GeV to 180 GeV. GEANT4 9.2 was used, comparing the predictions of different physics lists. The normalization uncertainty is 1%. (Courtesy of the Atlas Collaboration. Reproduction with permission)

the signal—is limited to short distances near the point of their generation during the typical trigger window of an LHC detector.

Comparing with the same Tilecal test beam, with the beam at 90 degrees with different physics lists of release 10.1, the version used for Run 2 simulation by ATLAS, demonstrates the change in physics performance from the revision of physics modeling in GEANT4. The energy resolution (Fig. 11.3), longitudinal shape (Fig. 11.4) and lateral shapes (Fig. 11.5) are compared with four physics lists that combine the QGS or Fritiof FTF string models with the Bertini or Binary cascade.

11.5.3 Background Estimation for CMS

Simulations were used to assess the required shielding for the CMS detector, to reduce the background from the interaction region p–p collisions [33] and the accelerator tunnel [114]. These employed a combination of tools: the STRUCT code was used for simulation the accelerator lattice and the scoring of particles lost at collimators, MARS to generate the particles entering the experimental area and FLUKA to model their interactions and fluxes in the detector and surrounding area. The study confirmed the need for shielding from the accelerator background and evaluated the proposed solutions. Key aspects were the impact on muon physics, together with the flexibility of optionally tracking the products of muon-

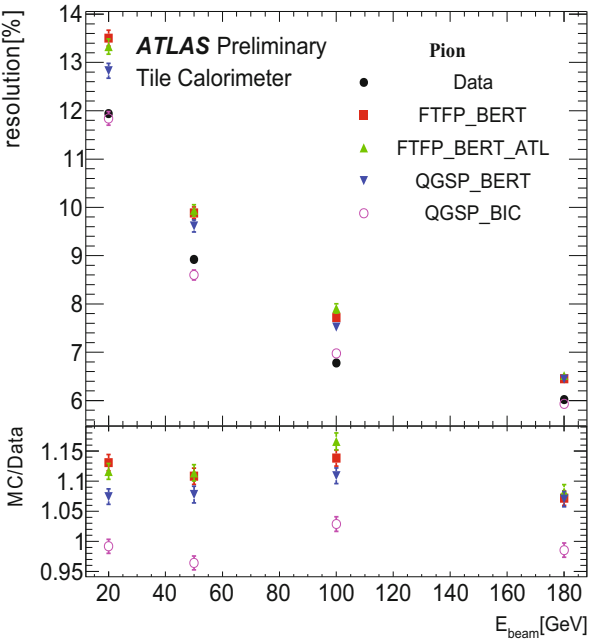


Fig. 11.3 Energy resolution of ATLAS Tilecal test beam compared to recent GEANT4 10.1 release, currently used in ATLAS simulation production. Lower panel is the ratio of simulated (MC) and data. Courtesy of the ATLAS collaboration (ATLAS public plot). The original data and comparisons of mean and RMS energy [113] were undertaken with GEANT4 version 9.2

nucleus interactions. The study identified high-energy muons as the most important remaining background, affecting, in particular, the innermost barrel chambers, contributing to approximately 10% to the total trigger rate (which is very small).

11.5.4 Validation from Comparisons with In-Situ Data

The accuracy of the simulation is determined using selected in-situ data from collisions, whenever possible. For example, ATLAS selected proton–proton collisions at 7 TeV (2010) and 8 TeV (2012) [115] to compare the energy deposition of charged hadrons with energies up to 30 GeV.

Decays of Λ, anti-Λ and K^0_s were used to identify π^+, π^-, protons and antiprotons and the ratio of their measured energy and momenta were compared with simulation using GEANT4 version 9.4. The ratio of Λ and anti-Λ to K^0_s is 40% higher in data than in simulation, but the normalized distributions are well reproduced within statistical precision, as seen in Fig. 11.6. The tail beyond $E/p > 1$ is due to the neutral background. The fraction of $E/p \leq 0$ is due to interactions before the calorimeter; it is underestimated about 10% by the simulation across all particle species. The difference is taken between particle species of the mean values of E/p in order to reduce the effect of the neutral background. The difference between π^- and antiprotons, due to extra energy from the antiprotons annihilation, is described within uncertainties by the FTFP_BERT physics list in GEANT4 version 9.4.

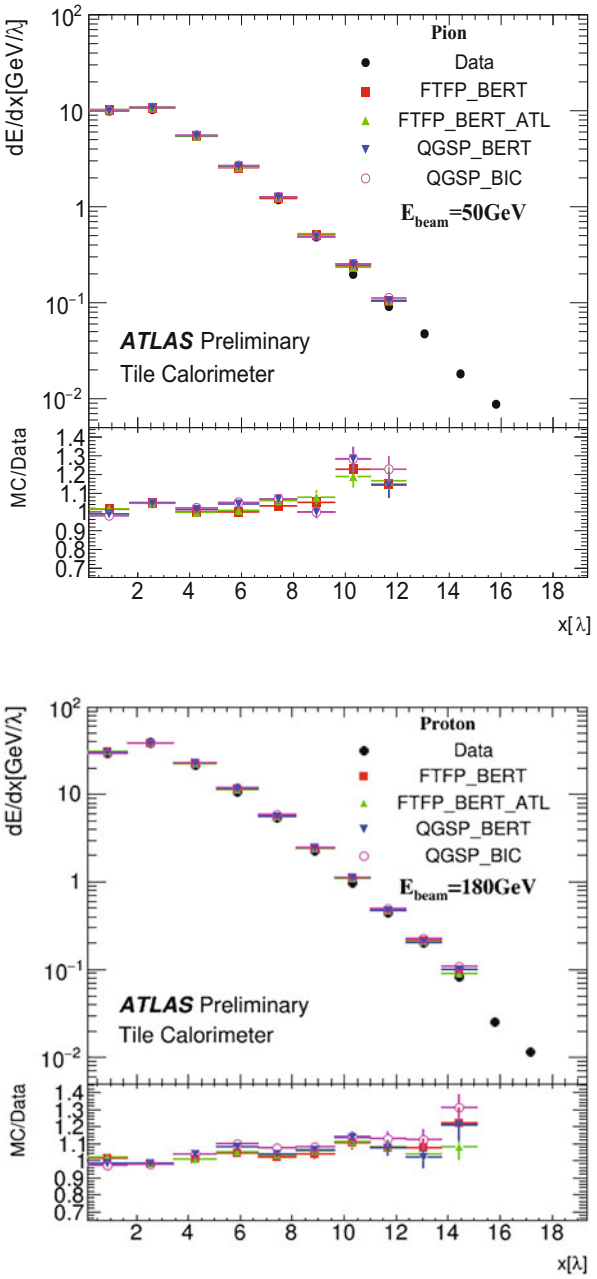


Fig. 11.4 Longitudinal shower profile of 50 GeV pion and 180 GeV proton beams in ATLAS Tilecal, with the modules placed at 90 degrees in dedicated test beam setup, versus recent GEANT4 10.1. Lower panel is the ratio of simulated (MC) and data. Courtesy of the ATLAS collaboration (ATLAS public plot). The original data and earlier comparisons (vs. GEANT4 ver. 9.2) were in Ref. [113]

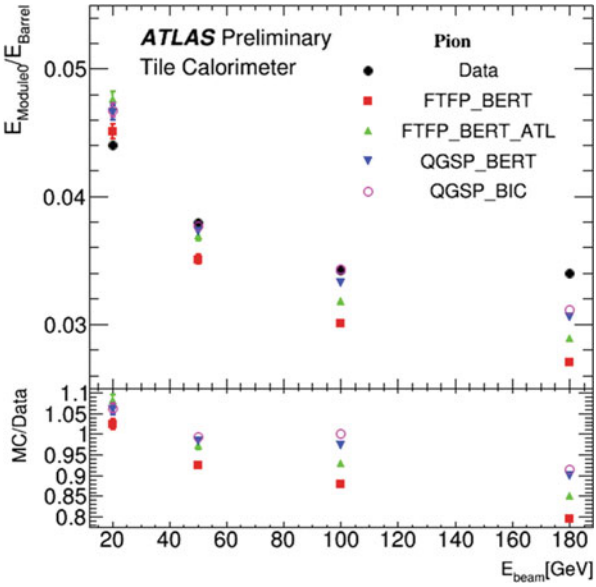


Fig. 11.5 Lateral spread of 20, 50, 100 and 180 GeV pions incident on the ATLAS Tile Calorimeter at 90-degree angle. Black points represent data obtained in the period 2000–2003, and the colored points simulations using different physics lists of GEANT4 version 10.1. Lower panel is ratio of simulated (MC) and data. Courtesy of the ATLAS collaboration (ATLAS public plot). The original publication [113] compared data with the earlier version 9.2 of GEANT4

Inclusive spectra of isolated hadron tracks were used to compare *E/p* distributions with simulation. After the energy deposition of neutral particles is subtracted, a 5% discrepancy was found in the response to isolated charged hadrons between the modelling using two GEANT4 hadronic physics lists (FTFP_BERT and QGSP_BERT) in the central region of the calorimeter (Fig. 11.7).

When tracks that interact only in the hadron calorimeter were examined separately, the detector simulation was found to describe the response well. Tracks that interacted only in the EM calorimeter showed discrepancies 5–10% in E/p, which is consistent with being the origin of the difference of all tracks.

These comparisons are used in one of the methods of estimating the uncertainty of the jet energy scale. Compared with the most recent estimations of the jet energy scale, these estimates have larger uncertainties over most of the energy range, but confirm estimates from in situ beam data. However, they currently provide the only estimate for the largest momenta ($p_T > 2$ TeV).

11.5.5 The Estimation of Jet Energy Scale in ATLAS and CMS

The earliest estimates of the jet energy scale, in the first years of the operation of ATLAS and CMS relied critically on detector simulation.

In CMS the detector simulation using GEANT4 was used in multiple stages of the initial calibration of the jet energy scale [116]. Initially, it was used to determine a base calibration factor $C_{MC}(p_T^{reco})$ to account for the fraction of jet energy not

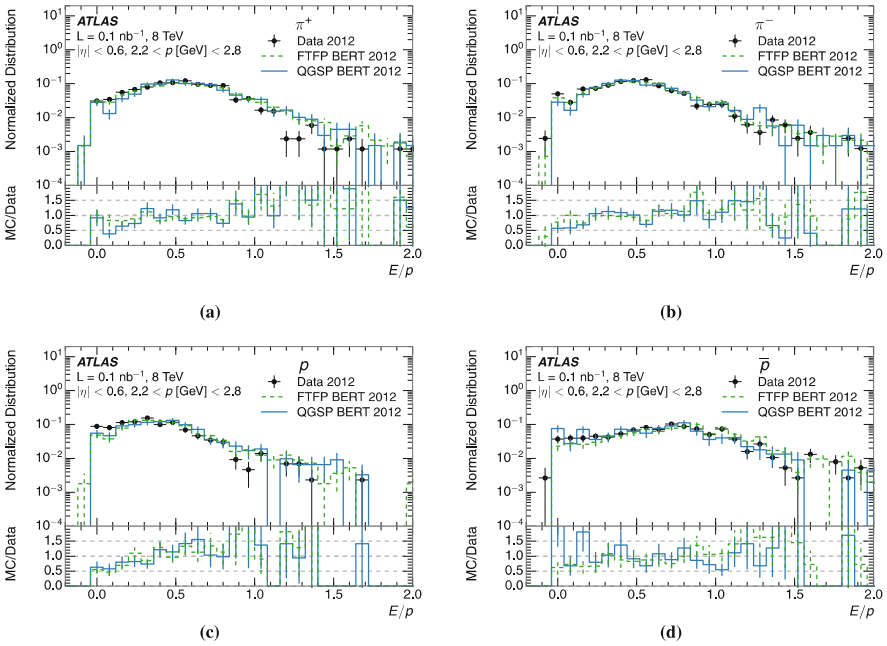


Fig. 11.6 The E/p distribution for (**a**) π^+, (**b**) π^-, (**c**) protons and (**d**) antiprotons from selected ATLAS 8 TeV data of identified Λ, anti-Λ and Ks decays [115] with $|\eta| < 0.6$ and $2.2 < p/\text{GeV}/2.8$. The lower part of each panel shows the ratio of MC simulation (using GEANT4 ver. 9.4) to data. Reproduced from [115] under the Creative Commons License 4.0

observed due to the inactive parts, and to determine the variation of response for different types of particles.

Subsequently, the balance of the transverse momenta of γ + jets and Z + jets events was used to compare the well-measured electromagnetic response to the hadron/jet response predicted by simulation for different p_T and η values. CMS concluded that the Monte Carlo correction factor described the bulk of non-uniformity of C_{MC} in η and non-linearity in p_T. The estimates of the data/MC ratio of jet energy using the different samples (Fig. 11.8) were consistent, flat in p_T and its value was $R_{data/MC} = 0.985 \pm 0.001$.

The estimation of the jet energy scale with Run II data relies less on simulation. Comparisons versus simulation based on a newer GEANT4 version (10.2) found that a larger correction was required, partially ascribed to the migration to Fritiof-based hadronic models.

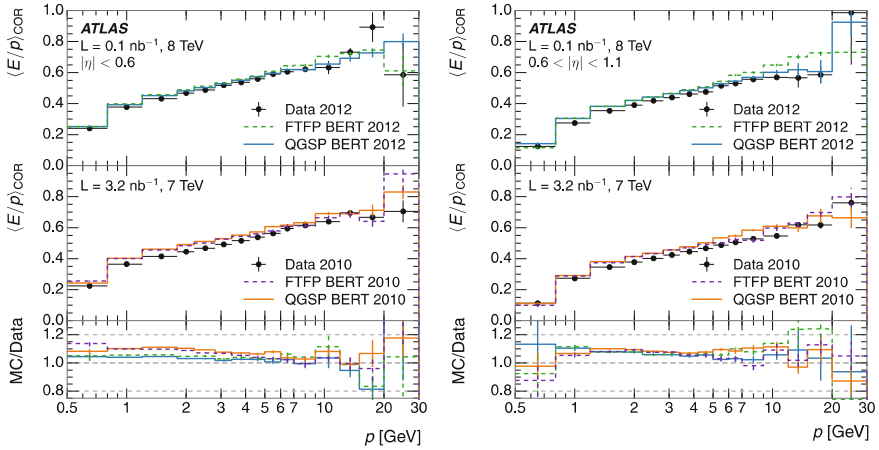


Fig. 11.7 Comparison of the average ratio <E/p>$_{COR}$ of charged tracks of cluster energy in ATLAS calorimeters versus momentum p measured in the tracker. Comparisons within (**a**) $|\eta| < 0.6$, and (**b**) $0.6 < |\eta| < 1.1$, obtained after subtraction of corresponding estimates of neutral particle response, versus track momentum. Tracks with no matching energy cluster in the calorimeter are included. The bottom portion of the panels shows the ratio of simulation (using GEANT4 version 9.4 and two physics lists) to data. Error bars are statistical. Reproduced from [115] under the Creative Commons License 4.0

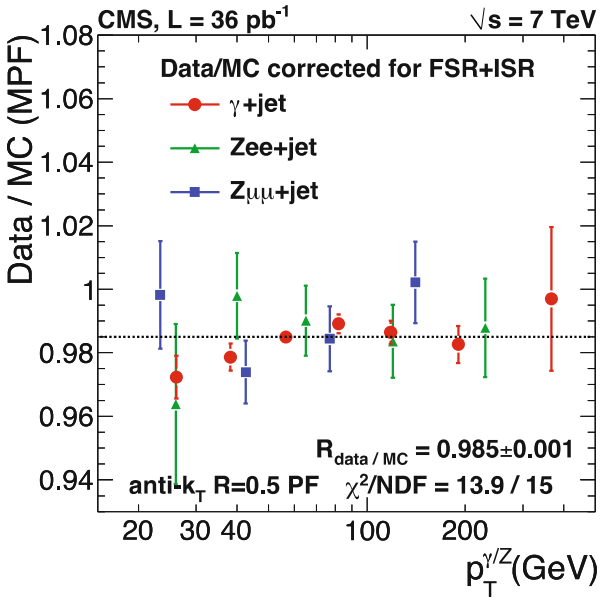


Fig. 11.8 Correction factor from comparison of γ/W + jet events in CMS 2010 data at $\sqrt{s} = 7$ TeV. From Ref. [116], reproduced under the Creative Commons license 3.0

11.5.6 Fast Simulation in CMS During LHC Run I and II

The fast simulations of the LHC experiments are sophisticated programs, combining the most important physics processes for electrons, gammas in the trackers and muons in the muon systems, with sampling from parameterized distributions for the showering of electrons, gammas and hadrons in the calorimeters. CMS's fast simulation FastSim [117] is a simplified geometrical description of the tracker, refined to obtain percent-level agreement for photon conversion, using around 30 thin nested cylinders.

The fast simulation in the tracker reconstructs each track using only its own generated hits, and cannot reproduce fake tracks that result from the incorrect association of hits. Though good agreement is seen in comparisons with Run-I data, the limitations of this approach are apparent in modeling the efficiency of track reconstruction and the fake rate [97].

Electron and hadron showers in calorimeters are turned into energy spot hits directly, distributed according to a Γ-function with parameters which fluctuate between showers using GLASH [118] or a similar approach.

Regular comparisons with the full detector simulation are used to monitor all the quantities used in physics analyses [119]. Agreement is observed at a level of 10%. Good agreement is particularly important for the missing transverse energy E_T.

During Run-I the CMS FastSim was used for the parameter scans for SuperSymmetry searches and samples of events used to evaluate systematic uncertainties [97], because the computing resources required for full simulation would have exceeded the available ones.

Once in-situ data is available, they are used as the final yardstick of the quality of both the fast and full simulation, and are used to address possible discrepancies between fast and detailed simulation.

11.5.7 Future Detectors: Fine-Grained Calorimeters of CALICE

Proposed detectors for next-generation collider experiments rely on particle flow reconstruction methods to obtain the required energy resolution for their physics programme. In order to obtain this performance, it is necessary to accurately model the energy deposition of charged hadrons, in order to subtract them from the observed signals.

Measurements with fine-grained calorimeters provide the most promising methods to validate the most important properties of detector simulation tools, and in particular their hadronic modeling. The CALICE experiment has undertaken test beam measurements of prototype calorimeters with many layers of scintillator tiles of fine granularity.

The prototype calorimeter with analog readout consisted of 38 layers, each containing a steel absorber plate and a scintillator layer. The 30×30 cm^2 core of the scintillator had granularity 3×3 cm^2, and outer regions 6×6 cm^2 and 12×12 cm^2. Data was collected with pions between 8 GeV and 100 GeV and compared with Monte Carlo simulation [120] using GEANT4 version 9.4.

The fine segmentation allows the estimation of the layer of the first hard interaction. This is used to obtain the shower profile relative to this starting layer. Averaging over showers starting in different layers reduces the effect of the variation of calibration, and is used to estimate uncertainties.

Comparisons of the mean longitudinal and lateral shower profiles for 8 GeV, 18 GeV and 80 GeV pions with GEANT4 physics lists including QGSP_BERT and FTFP_BERT were provided.

Figure 11.9 top row shows the longitudinal shower profile for pions of 8 GeV (left), 18 GeV (center) and 80 GeV (right) compared with GEANT4 physics lists FTFP_BERT.

Normalization to unit total is used in each distribution. For FTFP_BERT less energy is deposited in the early shower layers at all energies. At 80 GeV a difference is seen in the shower maximum of 10% (FTF versions) to 20% (QSGP versions), and the shower is more compact.

Similar comparisons for the radial shower profile show that all physics lists underestimate the radial extent of the showers and have a larger fraction of energy in the core. The effect is most pronounced at 80 GeV, see Fig. 11.10.

Either improvement of the relevant models or alternative physics models is needed to better describe these shower shapes, and provide the accuracy to use the full potential of simulation for future highly granular calorimeters.

11.6 Applications in Other Fields

Particle transport simulation tools, including GEANT4 and FLUKA, have seen greatly increasing usage beyond High Energy and Nuclear Physics (HENP) experiments in the past decade.

In particular in medical physics, their application has seen spectacular growth, and has spanned several domains, especially the development and refinement of new methods and assessment of treatments in radiotherapy, and the simulation of medical imaging detectors.

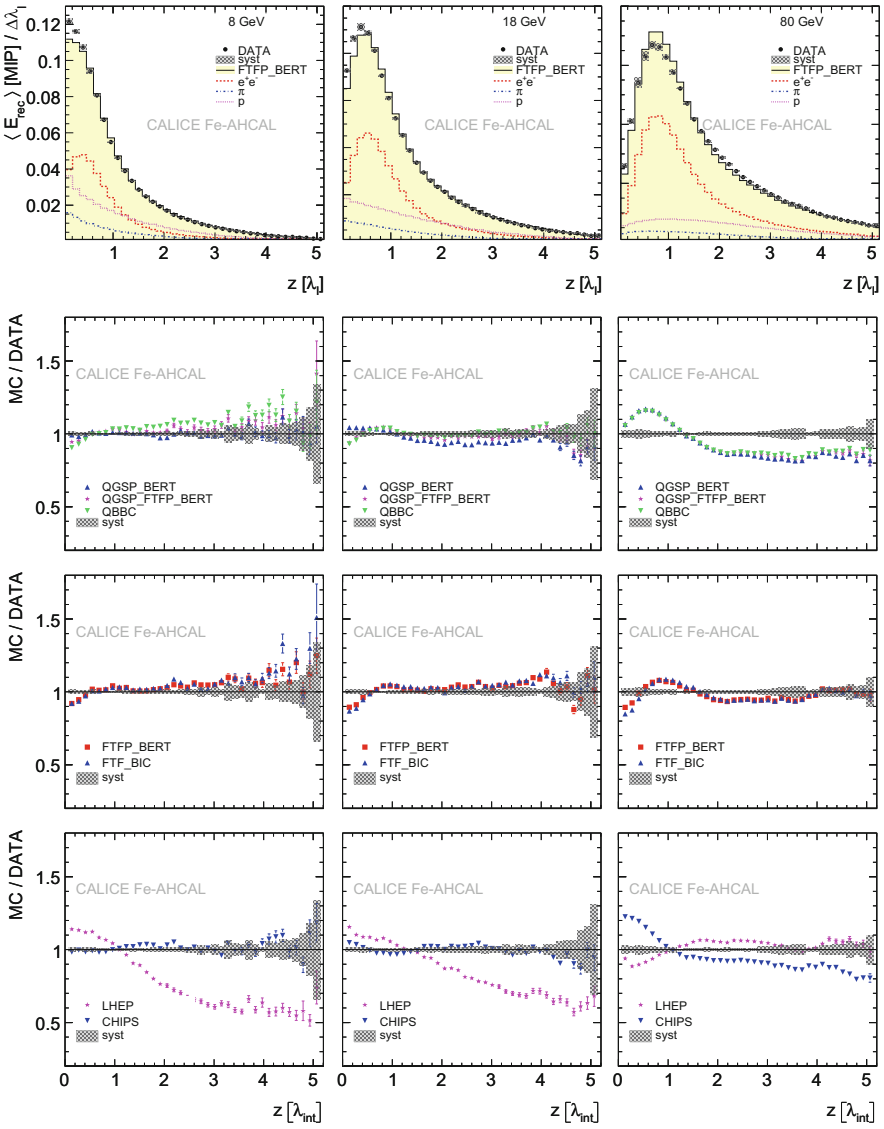


Fig. 11.9 Mean longitudinal shower profile, starting at the layer of the first interaction, for pion beams in CALICE iron-scintillator analog hadronic calorimeter. Pion energies are 8 GeV (left), 18 GeV (center) Top row: data (circles) compared with FTFP_BERT physics list of GEANT4 version 9.4. The parts deposited by different particles (electrons/positrons, pions and protons) in the simulation are shown. Lower 3 rows: ratios between selected physics lists and data. (We note that the CHIPS and LHEP physics lists were withdrawn in Geant4 release 10.0). Reproduced from [120], under Creative Commons License 3.0

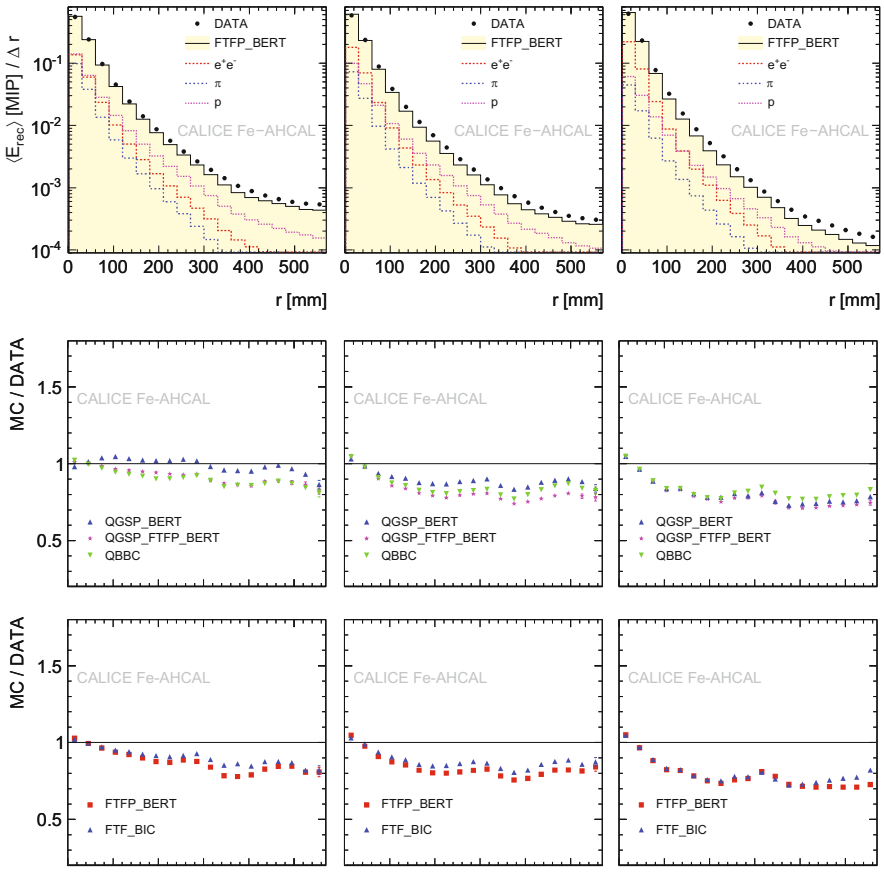


Fig. 11.10 Mean radial shower profile for pion beams in CALICE iron-scintillator analog hadronic calorimeter. Pion energies are 8 GeV (left), 18 GeV (center) and 80 GeV (right). Top row: data (circles) compared with FTFP_BERT physics list of GEANT4 version 9.4. Deposits by different particles (electrons/positrons, pions and protons) in the simulation are shown. Centre and lower row: ratios between selected physics lists and data, including QGSP_BERT and FTFP_BERT and variants. Reproduced from [120], under Creative Commons License 3.0. (Physics lists using models withdrawn in subsequent releases, CHIPS and LHEP, are omitted)

The simulation tools are used widely also in determining the effect of radiation on satellites and spacecraft from planetary radiation environments and the solar and galactic rays. Specialized tools have been developed for shielding studies [121] as have general purpose tools to evaluate effects of the space environments [122].

FLUKA is used for shielding and target design of accelerators, activation studies, and also for cosmic ray studies, due to its ability to simulate up to 20 PeV. A further application is the assessment of dose to aircrews flying in commercial aircraft.

11.6.1 Medical Imaging

A key application of Monte Carlo simulation in medical imaging is the development of novel instrumentation, e.g. progress in Position Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT). Particle transport simulation enables an evaluation of new materials, geometries or system configurations in multiple versions without the expense of always creating a hardware prototype [123].

Dedicated particle transport tools have been developed specifically for the simulation of PET or SPECT devices. An early tool PET-EGS [124] used EGS, with GEANT4 used in the leading tool, GATE [125], and in GAMOS [126], and Penelope in PeneloPET [127].

GATE is one of the most commonly used dedicated simulation tools for PET [123]. This is due to its simplicity in generating setups and steering the simulation using text commands, and due to the benefits of the validation of the GEANT4 toolkit.

Particle transport simulation is a standard tool in many applications. However, one of its key drawbacks is the large computation times required. Methods have been therefore developed to mitigate this problem. As an example, networked computers are being employed to speed up the calculations. Alternatively, hybrid computational models are being employed, such as generating the initial photons using SIMSET [128] or using EGSnrc as its core simulation engine [129].

11.6.2 Proton and Hadron Therapy

As in photon radiotherapy, the recent advances of proton and ion beam therapy have heavily relied on radiation transport Monte Carlo tools [130]. Due to the need for short computation time, specialized Treatment Planning Systems (TPS) with analytical or simplified models for the fast estimation of dose delivery are the clinical standard.

GEANT4 validation for proton-therapy involves selection of the best performing physics models [131]. These tools are also essential in evaluating potential improvements in TPS methods [132].

Another critical aspect is the simulation of the effects of organ motion on dose delivery, e.g. with the GEANT4-based GATE and standalone applications [133].

Specialized applications were developed to use GEANT4 in particle therapy and provide tailored and validated physics configurations, interfaces to CT input, and tools including the reading and writing of snapshots of particles at specific interfaces as phase space files. In the past decade two dedicated applications, PTSim [134] and TOPAS [135] targeted easy use by clinical physicists in Japan and the US respectively. TOPAS emphasized reliable configuration, and modelling of the motion

of components. Both have seen increasing use for research and in selected clinical settings.

The therapy potential of light ion beams has been the topic of increasing research during the past decade. A key drawback is the energy deposition of nuclear fragments which extends beyond the Bragg peak. Early comparisons identified discrepancies of some tens of percent in non-differential quantities between data and MC [136].

In more recent studies better agreement in these tail dose depositions and dose profiles was obtained with GEANT4 and FLUKA, but differences in prompt gamma emission continue to be an issue [137]. The lateral beam widening is also well confirmed by FLUKA results [138].

A key need in ion therapy is the monitoring of the range of ions. The detection of positron emission in a PET detector has been in clinical use, and a newer method involves prompt gamma emission. Both techniques have been investigated using particle transport to quantify the location of emission and the spectra of clinically interesting gammas.

Modeling of ion–ion interactions at therapeutic energies is frequently at the edge of applicability for cascade models (below 150–200 MeV), and the resulting spectra are influenced by the details of many nuclear de-excitation processes. Discrepancies in secondary particle production in FLUKA were improved with the addition of a Boltzman Master Equation (BME) model and other modelling refinements [139]. New measurements have been made with ion projectiles to provide data for comparison and improvement of modelling. One set using lower energy (62 A MeV) ^{12}C beam measured a large set of secondary spectra (p, d, t through to ^{11}B) [140]. Comparison with GEANT4 models revealed the need for improvement of the modeling used (binary cascade and QMD).

FLUKA's existing applications in particle therapy [141] include the production of data for Treatment Planning Systems (TPS), checking the plans created by TPS in selected cases for quality control and improvement of patient dose delivery, and in feasibility and sensitivity studies of prompt gammas for range and dose monitoring. Another use has been the monitoring of dose delivery in ion therapy through PET imaging of positron emitter production, undertaken either after treatment or through an integrated PET device during the patient treatment.

11.6.3 Developments for Microdosimetry and Nanodosimetry

New models and adaptation of physics models have been developed to extend their application to smaller energies. Dedicated Monte Carlo track structure codes have been used in the investigation of radiation effects at the micron scale and at scales appropriate for biological research [142] and modeling of radiotherapy outcomes [143]. GEANT4 has been extended to provide track structure modeling in liquid and gaseous water with the development of the GEANT4-DNA package [144]. This has enabled its use in many applications in these fields [145, 146].

The GEANT4-DNA package provides new physical models for the description of elastic and inelastic electromagnetic interactions of electrons and select ions (Li to O, plus Si and Fe) in liquid water, previously only available in dedicated 'track structure' codes such as PARTRAC [147]. In addition, GEANT4-DNA offers features to model the water radiolysis from ionizing radiation: ionized or excited water molecules and water anions are generated and tracked at a physicochemical stage up to a few picoseconds, and subsequently in a 'chemical' stage using models of generation, diffusion and reaction of new chemical species. This is part of an effort to model and understand the first stages of DNA damage.

11.7 Outlook

During the past decade the application of detector simulation tools has been significantly widened through the implementation of improved physics models. Code and models have become more accurate in describing benchmark data. The need for more accurate data for comparison and model improvement has been one motivation for some thin-target (HARP, MIPP) and thick target (CALICE) experiments.

The requirements that arise from the projected use of simulations as an integral part of the next generation detectors becomes ever stronger. Witness, e.g. the use of particle flow reconstruction [148] in proposed experiments at the energy frontier, including the Linear Collider and the Future Circular Collider, to address one of their major challenges.

The need for further development of physics models for high-energy hadron–nucleus interactions is evident. Several promising approaches are being pursued, including the extension and tuning of existing implementations of current models (Fritiof, Quark-Gluon String), the incorporation of alternative implementations of existing models, such as QGSJET, complementary modeling approaches (DPMJET) and the incorporation of new models (EPOS). The availability of high quality thin target experimental data at energies over a range of momenta from 20 to 158 GeV/c [149] is an important resource; lack of data for higher energies is a constraint.

In addition to physics improvements, the large increase in statistics of simulated events for the HL-LHC requires a large improvement in CPU performance, of approximately a factor of ten. Research in CPU-performance and emerging architectures in the GeantV R&D effort indicate a more realistic target of a factor of 2–4 in performance improvement may be within reach for detailed simulation.

These prospects strengthen the need for parameterized (fast) simulation methods which can reproduce the results of detailed simulation as accurately as possible, for use in a majority and potentially an ever-larger fraction of analyses. Hybrid methods combining parameterized and detailed simulation in innovative ways and the machine learning approach to parameterized/fast simulation appear amongst the options which will see significant development and research in the next years.

Bibliography

1. Graf, N., McCormick, J.: Simulator For The Linear Collider (SLIC): A Tool For ILC Detector Simulations. In: AIP Conference Proceedings. pp. 503–512. AIP (2006)
2. Frank, M., Gaede, F., Nikiforou, N. et al.: DDG4 A Simulation Framework based on the DD4hep Detector Description Toolkit. J. Phys. Conf. Ser. 664, 072017 (2015). https://doi.org/10.1088/1742-6596/664/7/072017
3. Behnke, T., Brau, J.E., Burrows, P.N. et al.: The International Linear Collider Technical Design Report - Volume 4: Detectors. Batavia, IL (United States) (2013)
4. Daniel Elvira, V.: Impact of detector simulation in particle physics collider experiments. Phys. Rep. 695, 1–54 (2017). https://doi.org/10.1016/J.PHYSREP.2017.06.002
5. Albrecht, J., Alves, A.A., Amadio, G. et al.: A Roadmap for HEP Software and Computing R&D for the 2020s. (2017). https://doi.org/10.1007/s41781-018-0018-8
6. Alves, A.A., Amadio, G., Anh-Ky, N. et al.: A Roadmap for HEP Software and Computing R&D for the 2020s. (2017)
7. Amadio, G., Apostolakis, J., Bandieramonte, M. et al.: The GeantV project: Preparing the future of simulation. J. Phys. Conf. Ser. 664, (2015). https://doi.org/10.1088/1742-6596/664/7/072006
8. Sjöstrand, T., Ask, S., Christiansen, J.R. et al.: An introduction to PYTHIA 8.2. Comput. Phys. Commun. 191, 159–177 (2015). https://doi.org/10.1016/j.cpc.2015.01.024
9. Andersson, B., Gustafson, G., Pi, H.: The FRITIOF model for very high energy hadronic collisions. Zeitschrift fr Phys. C Part. Fields. 57, 485–494 (1993). https://doi.org/10.1007/BF01474343
10. Sjöstrand, T., Bengtsson, M.: The Lund Monte Carlo for jet fragmentation and e+ e- physics - jetset version 6.3 - an update. Comput. Phys. Commun. 43, 367–379 (1987). https://doi.org/10.1016/0010-4655(87)90054-3
11. Corcella, G., Knowles, I.G., Marchesini, G. et al.: HERWIG 6: an event generator for hadron emission reactions with interfering gluons (including supersymmetric processes). J. High Energy Phys. 2001, 10 (2001)
12. Bellm, J., Gieseke, S., Grellscheid, D. et al.: Herwig 7.0/Herwig++ 3.0 release note. Eur. Phys. J. C. 76, 196 (2016). https://doi.org/10.1140/epjc/s10052-016-4018-8
13. Agostinelli, S., Allison, J., Amako, K. et al.: Geant4—a simulation toolkit. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 506, 250–303 (2003). https://doi.org/10.1016/S0168-9002(03)01368-8
14. Nakao, N., Mokhov, N.V: MARS15 code in accelerator applications. http://www-ap.fnal.gov/users/mokhov/papers/2007/Conf-07-416-APC.pdf (2007)
15. Brun, R., Giani, S.: GEANT—Detector description and simulation tool. (1994)
16. Fesefeldt, H.S.: Simulation of hadronic showers, physics and applications. Physikalisches Institut, RWTH Aachen Physikzentrum, 5100 Aachen, Germany (1985)
17. Zeitnitz, C., Gabriel, T.A.: The GEANT-CALOR interface and benchmark calculations of ZEUS test calorimeters. Nucl. Instr. Methods A. 349, 106–111 (1994). https://doi.org/10.1016/0168-9002(94)90613-0
18. Gabriel, T.A., Bishop, B.L., Brau, J.E.: The physics of compensating calorimetry and the new calor89 code system. IEEE Trans. Nucl. Sci. 36, 14–22 (1989). https://doi.org/10.1109/23.34394
19. Fassò, A., Ferrari, A., Ranft, J., et al.: FLUKA: present status and future developments. In: Menzione, A. and Scribano, A.P.G. 493 (eds.) Proc. IV Int. Conf. on Calorimetry in High Energy Physics, La Biodola, Italy, 21-26 Sept. 1993. p. World Scientific. World Scientific (1993)
20. Battistoni, G., Boehlen, T., Cerutti, F. et al.: Overview of the FLUKA code. Ann. Nucl. Energy. 82, 10–18 (2015). https://doi.org/10.1016/J.ANUCENE.2014.11.007

21. Roesler, S., Engel, R., Ranft, J.: The Monte Carlo Event Generator DPMJET-III. In: A. Kling M. Nakagawa L. Távora & P. Vaz PG - 1033, F.B. (ed.) Advanced Monte Carlo for Radiation Physics, Particle Transport Simulation and Applications. p. 1038 (2001)
22. Allison, J., Amako, K., Apostolakis, J. et al.: Recent developments in Geant4. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 835, 186–225 (2016). https://doi.org/10.1016/j.nima.2016.06.125
23. Sihver, L., Sato, T., Gustafsson, K. et al.: Iwase, H., Niita, K., Nakashima, H., Sakamoto, Y., Iwamoto, Y., Matsuda, N.: An update about recent developments of the PHITS code. Adv. Sp. Res. 45, 892–899 (2010). https://doi.org/10.1016/j.asr.2010.01.002
24. Mashnik, S.G.: Validation and Verification of MCNP6 Against Intermediate and High-Energy Experimental Data and Results by Other Codes. Eur. Phys. J. Plus. (2011). https://doi.org/10.1140/epjp/i2011-11049-1
25. Goorley, T., James, M., Booth, T. et al.: Features of MCNP6. Ann. Nucl. Energy. 87, 772–783 (2016). https://doi.org/10.1016/J.ANUCENE.2015.02.020
26. Veenhof, R.: Garfield, recent developments. Nucl. Instruments Methods Phys. Res. A. 419, 726–730 (1998). https://doi.org/10.1016/S0168-9002(98)00851-1
27. Smirnov, I.B.: Modeling of ionization produced by fast charged particles in gases. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 554, 474–493 (2005). https://doi.org/10.1016/j.nima.2005.08.064
28. Veenhoff, R., Schindler, H.: Garfield++ – simulation of ionisation based tracking detectors, http://garfieldpp.web.cern.ch/garfieldpp/
29. Pfeiffer, D., De Keukeleere, L., Azevedo, C. et al.: A Geant4/Garfield++ and Geant4/Degrad Interface for the Simulation of Gaseous Detectors. (2018)
30. van der Ende, B.M., Rand, E.T., Erlandson, A. et al.: Use of SRIM and Garfield with Geant4 for the characterization of a hybrid 10B/3He neutron detector. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 894, 138–144 (2018). https://doi.org/10.1016/J.NIMA.2018.03.056
31. Lukas, W.: Fast Simulation for ATLAS: Atlfast-II and ISF. J. Phys. Conf. Ser. 396, 022031 (2012). https://doi.org/10.1088/1742-6596/396/2/022031
32. Mokhov, N.V, Gudima, K.K., James, C.C. et al.: Recent enhancements to the MARS15 code. Radiat Prot Dosim. 116, 99–103 (2005)
33. Huhtinen, M., Aarnio, P.A.: Neutron and photon fluxes and shielding alternatives for the CMS detector at LHC. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 363, 545–556 (1995). https://doi.org/10.1016/0168-9002(95)00444-0
34. Hrivnacova, I., Adamova, D., Berejnoi, V. et al.: The Virtual Monte Carlo. In: Computing in High Energy and Nuclear Physics (CHEP03), La Jolla, CA. p. arXiv:cs/0306005 (2003)
35. Kalos, M.H., Whitlock, P.A.: Monte Carlo methods. Wiley - VCH (2008)
36. James, F.: Monte Carlo theory and practice. Reports Prog. Phys. 43, 1145–1189 (1980). https://doi.org/10.1088/0034-4885/43/9/002
37. Bielajew, A.F.: Fundamentals of the Monte Carlo method for neutral and charged particle transport. (2000)
38. Panneton, F., L'Ecuyer, P.: Resolution-stationary random number generators. Math. Comput. Simul. 80, 1096–1103 (2010). https://doi.org/10.1016/j.matcom.2007.09.014
39. Savvidy, G., Ter-Arutyunyan-Savvidy, N.: On the Monte Carlo simulation of physical systems. J. Comput. Phys. 97, 566–572 (1991). https://doi.org/10.1016/0021-9991(91)90015-D
40. Savvidy, K., Savvidy, G.: Spectrum and entropy of C-systems MIXMAX random number generator. Chaos, Solitons & Fractals. 91, 33–38 (2016). https://doi.org/10.1016/j.chaos.2016.05.003
41. Lüscher, M.: A portable high-quality random number generator for lattice field theory simulations. Comput. Phys. Commun. 79, 100–110 (1994). https://doi.org/10.1016/0010-4655(94)90232-1
42. James, F.: RANLUX: A Fortran implementation of the high-quality pseudorandom number generator of Lüscher. Comput. Phys. Commun. 79, 111–114 (1994). https://doi.org/10.1016/0010-4655(94)90233-X

43. L'Ecuyer, P., Simard, R.: TestU01: A C library for empirical testing of random number generators. ACM Trans. Math. Softw. 33, 22–es (2007). https://doi.org/10.1145/1268776.1268777
44. Berger, M.J.: Monte Carlo calculation of the penetration and diffusion of fast charged particles. In: B. Alder S. Fernbach and Rotenberg, M. (eds.) Methods in Computational Physics: Advances in Research and Applications, Vol. 1. Statistical Physics. pp. 135–215. Academic, New York (1963)
45. Ford, R.L., Nelson, W.R.: The EGS Code System: Computer Programs for the Monte Carlo Simulation of Electromagnetic Cascade Showers (Version 3). (1978)
46. Nelson, W.R., Hirayama, H., Rogers, D.W.O.: The EGS4 code system. (1985)
47. Salvat, F., Fernández-Varea, J.M.: Overview of physical interaction models for photon and electron transport used in Monte Carlo codes. Metrologia. 46, S112–S138 (2009). https://doi.org/10.1088/0026-1394/46/2/S08
48. Baró, J., Sempau, J., Salvat, F. et al.: PENELOPE: An algorithm for Monte Carlo simulation of the penetration and energy loss of electrons and positrons in matter. Nucl. Instruments Methods Phys. Res. B. 100, 31–46 (1995). https://doi.org/10.1016/0168-583X(95)00349-5
49. Goudsmit, S., Saunderson, J.L.: Multiple Scattering of Electrons. II. Phys. Rev. 58, 36–42 (1940). https://doi.org/10.1103/PhysRev.58.36
50. Molière, G.: Theorie der Streuung schneller geladener Teilchen I. Einzelstreuung am abgeschirmten Coulomb-Feld. Zeitschrift Naturforsch. Tl. A. 2, 133–+ (1947)
51. Lewis, H.W.: Multiple Scattering in an Infinite Medium. Phys. Rev. 78, 526–529 (1950). https://doi.org/10.1103/PhysRev.78.526
52. Bielajew, A.F., Rogers, D.W.O.: Presta: The parameter reduced electron-step transport algorithm for electron monte carlo transport. Nucl. Instruments Methods Phys. Res. Sect. B Beam Interact. with Mater. Atoms. 18, 165–171, 174–181 (1986). https://doi.org/10.1016/S0168-583X(86)80027-1
53. Kawrakow, I., Bielajew, A.F.: On the condensed history technique for electron transport. Nucl. Instruments Methods Phys. Res. B. 142, 253–280 (1998). https://doi.org/10.1016/S0168-583X(98)00274-2
54. Bielajew, A.F., Salvat, F.: Improved electron transport mechanics in the PENELOPE Monte-Carlo model. Nucl. Instruments Methods Phys. Res. B. 173, 332–343 (2001). https://doi.org/10.1016/S0168-583X(00)00363-3
55. Salvat, F., Fernández-Varea, J.M., Sempau, J.: "PENELOPE, A Code System for Monte Carlo Simulation of Electron and Photon Transport.", Barcelona (2009)
56. Kawrakow, I.: Accurate condensed history Monte Carlo simulation of electron transport. I. EGSnrc, the new EGS4 version. Med. Phys. 27, 485–498 (2000)
57. Hirayama, H., Namito, Y., Bielajew, A.F. et al.: The EGS5 code system. (2005)
58. Vilches, M., García-Pareja, S., Guerrero, R. et al.: Monte Carlo simulation of the electron transport through thin slabs: A comparative study of penelope, geant3, geant4, egsnrc and mcnpx. Nucl. Instruments Methods Phys. Res. Sect. B Beam Interact. with Mater. Atoms. 254, 219–230 (2007). https://doi.org/10.1016/j.nimb.2006.11.061
59. Incerti, S., Ivanchenko, V., Novak, M.: Recent progress of Geant4 electromagnetic physics for calorimeter simulation. J. Instrum. 13, C02054–C02054 (2018). https://doi.org/10.1088/1748-0221/13/02/C02054
60. Grichine, V.M., Sadilov, S.S.: Geant4 models for X-ray transition radiation. Nucl. Instruments Methods Phys. Res. A. 522, 122–125 (2004). https://doi.org/10.1016/j.nima.2004.01.031
61. Armstrong, T.W., Chandler, K.G.: HETC - a high energy transport code. Nucl. Sci. Eng. 49, 110–111 (1972)
62. Dementyev, A. V, Sobolevsky, N. M.: SHIELD - Universal Monte Carlo Hadron Transport Code: Scope and Applications. Radiat. Meas. 50, 553–557 (1999). https://doi.org/10.1016/S1350-4487(99)00231-0
63. Weisskopf, V.F., Ewing, D.H.: On the Yield of Nuclear Reactions with Heavy Elements. Phys. Rev. 57, 472–485 (1940). https://doi.org/10.1103/PhysRev.57.472

64. Botvina, A.S., Iljinov, A.S., Mishustin, I.N. et al.: Statistical simulation of the break-up of highly excited nuclei. Nucl. Phys. A. 475, 663–686 (1987). https://doi.org/10.1016/0375-9474(87)90232-6
65. Furihata, S.: The GEM Code - the Generalized Evaporation Model and the Fission Model. In: A. Kling, Barão, F., Nakagawa, M., Távora, L., and P. Vaz (eds.) Advanced Monte Carlo for Radiation Physics, Particle Transport Simulation and Applications. p. 1045–+ (2001)
66. Gudima, K.K., Mashnik, S.G., Toneev, V.D.: Cascade-exciton model of nuclear reactions. Nucl. Phys. A. 401, 329–361 (1983). https://doi.org/10.1016/0375-9474(83)90532-8
67. Mashnik, S.G., Gudima, K.K., Prael, R.E. et al.: CEM03.03 and LAQGSM03.03 Event Generators for the MCNP6, MCNPX, and MARS15 Transport Codes. (2008). https://doi.org/10.1016/j.nimb.2010.09.005
68. Bertini, H.W.: Intranuclear-Cascade Calculation of the Secondary Nucleon Spectra from Nucleon-Nucleus Interactions in the Energy Range 340 to 2900 MeV and Comparisons with Experiment. Phys. Rev. 188, 1711–1730 (1969). https://doi.org/10.1103/PhysRev.188.1711
69. Bleicher, M., Zabrodin, E., Spieles, C. et al.: Relativistic hadron-hadron collisions in the ultra-relativistic quantum molecular dynamics model. J. Phys. G Nucl. Part. Phys. 25, 1859–1896 (1999)
70. Folger, G., Ivanchenko, V.N., Wellisch, J.P.: The Binary Cascade. Eur. Phys. J. A. 21, 407–417 (2004). https://doi.org/10.1140/epja/i2003-10219-7
71. Heikkinen, A., Stepanov, N., Wellisch, J.P.: Bertini intra-nuclear cascade implementation in Geant4. In: 13th Intern. Computing in High Energy and Nuclear Physics, (CHEP 2003): La Jolla, California, March 24–28, 2003. p. arXiv:nucl-th/0306008 (2003)
72. Duarte, H.: Particle production in nucleon induced reactions above 14 MeV with an intranuclear cascade model. Phys. Rev. C. 75, 24611 (2007). https://doi.org/10.1103/PhysRevC.75.024611
73. Degtyarenko, P.V., Kossov, M.V., Wellisch, H.-P.: Chiral invariant phase space event generator. Eur. Phys. J. A. 8, 217–222 (2000). https://doi.org/10.1007/s100500070108
74. Wright, D.H., Kelsey, M.H.: The Geant4 Bertini Cascade. Nucl. Instrum. Methods A. 804, 175–188 (2015). https://doi.org/10.1016/j.nima.2015.09.058
75. Boudard, A., Cugnon, J., Leray, S. et al.: Intranuclear cascade model for a comprehensive description of spallation reaction data. Phys. Rev. C. 66, 044615 (2002). https://doi.org/10.1103/PhysRevC.66.044615
76. Boudard, A., Cugnon, J., David, J.-C. et al.: New potentialities of the Liège intranuclear cascade model for reactions induced by nucleons and light charged particles. Phys. Rev. C. 87, 014606 (2013). https://doi.org/10.1103/PhysRevC.87.014606
77. Mancusi, D., Boudard, A., Cugnon, J. et al.: Extension of the Liège intranuclear-cascade model to reactions induced by light nuclei. Phys. Rev. C. 90, 054602 (2014). https://doi.org/10.1103/PhysRevC.90.054602
78. Napolitani, P., Schmidt, K.-H., Botvina, A.S. et al.: High-resolution velocity measurements on fully identified light nuclides produced in Fe 56 + hydrogen and Fe 56 + titanium systems. Phys. Rev. C. 70, 054607 (2004). https://doi.org/10.1103/PhysRevC.70.054607
79. Ricciardi, M. V., Armbruster, P., Benlliure, J. et al.: Light nuclides produced in the proton-induced spallation of U 238 at 1 GeV. Phys. Rev. C. 73, 014607 (2006). https://doi.org/10.1103/PhysRevC.73.014607
80. Leray, S., David, J.C., Khandaker, M. et al.: Results from the IAEA Benchmark of Spallation Models. J. Korean Phys. Soc. 59, 791 (2011). https://doi.org/10.3938/jkps.59.791
81. Akchurin, N., Bedeschi, F., Cardini, A. et al.: Lessons from Monte Carlo simulations of the performance of a dual-readout fiber calorimeter. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 762, 100–118 (2014). https://doi.org/10.1016/J.NIMA.2014.05.121
82. Capella, A., Sukhatme, U., Tan, C.-I. et al.: Dual parton model. Phys. Rep. 236, 225–329 (1994). https://doi.org/10.1016/0370-1573(94)90064-7

83. Kaidalov, A.B.: Interactions of hadrons and nuclei at superhigh energies and small-x physics. Nucl. Phys. B Proc. Suppl. 75, 81–88 (1999). https://doi.org/10.1016/S0920-5632(99)00218-2
84. Ostapchenko, S.: QGSJET-II: towards reliable description of very high energy hadronic interactions. Nucl. Phys. B Proc. Suppl. 151, 143–146 (2006). https://doi.org/10.1016/j.nuclphysbps.2005.07.026
85. Uzhinsky, V., Apostolakis, J., Galoyan, A. et al.: Antinucleus and nucleus cross sections implemented in Geant4. Phys. Lett. B. 705, 235–239 (2011). https://doi.org/10.1016/j.physletb.2011.10.010
86. Uzhinsky, V., Galoyan, A.: Effect of u u diquark suppression in proton splitting in Monte Carlo event generators. Phys. Rev. D. 91, 037501 (2015). https://doi.org/10.1103/PhysRevD.91.037501
87. Koning, A.J. et al.: The JEFF evaluated nuclear data project. In: O. Bersillon F. Gunsing, E.B.R.J. and S. Leray (eds.) International Conference on Nuclear Data for Science and Technology 2007 (April 22–27, 2007, Nice, France). pp. 194–199. EDP Sciences (2008)
88. Chadwick, M.B., Obložinský, P., Herman, M. et al.: ENDF/B-VII.0: Next Generation Evaluated Nuclear Data Library for Nuclear Science and Technology. Nucl. Data Sheets. 107, 2931–3060 (2006). https://doi.org/10.1016/j.nds.2006.11.001
89. Brown, D.A., Chadwick, M.B., Capote, R. et al.: ENDF/B-VIII.0: The 8thMajor Release of the Nuclear Reaction Data Library with CIELO-project Cross Sections, New Standards and Thermal Scattering Data. Nucl. Data Sheets. 148, 1–142 (2018). https://doi.org/10.1016/j.nds.2018.02.001
90. Shibata, K., Iwamoto, O., Nakagawa, T. et al.: JENDL-4.0: A New Library for Nuclear Science and Engineering. J. Nucl. Sci. Technol. 48, 1–30 (2011). https://doi.org/10.1080/18811248.2011.9711675
91. Ge, Z.G., Zhao, Z.X., Xia, H.H. et al.: The Updated Vversion of Chinese Evaluated Nuclear Data Library (CENDL-3.1). J. Korean Phys. Soc. 59, 1052–1056 (2011). https://doi.org/10.3938/jkps.59.1052
92. Kossov, M.V.: Chiral-invariant phase space model. Eur. Phys. J. A - Hadron. Nucl. 14, 265–269 (2002). https://doi.org/10.1140/epja/i2001-10211-3
93. Gabriel, T.A., Groom, D.E., Job, P.K. et al.: Energy dependence of hadronic activity. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 338, 336–347 (1994). https://doi.org/10.1016/0168-9002(94)91317-X
94. Wigmans, R.: Calorimetry : Energy measurement in particle physics. Clarendon Press (2000)
95. Groom, D.E.: Energy flow in a hadronic cascade: Application to hadron calorimetry. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 572, 633–653 (2007). https://doi.org/10.1016/j.nima.2006.11.070
96. de Favereau, J., Delaere, C., Demin, P. et al.: DELPHES 3: a modular framework for fast simulation of a generic collider experiment. J. High Energy Phys. 57 (2014). https://doi.org/10.1007/JHEP02(2014)057
97. Giammanco, A.: The Fast Simulation of the CMS Experiment. J. Phys. Conf. Ser. 513, 022012 (2014). https://doi.org/10.1088/1742-6596/53/2/022012
98. Ritsch, E., Collaboration, the A.: Concepts and Plans towards fast large scale Monte Carlo production for the ATLAS Experiment. J. Phys. Conf. Ser. 523, 012035 (2014). https://doi.org/10.1088/1742-6596/523/1/012035
99. Paganini, M., de Oliveira, L., Nachman, B.: CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. Phys. Rev. D. 97, 014021 (2018). https://doi.org/10.1103/PhysRevD.97.014021
100. Wigmans, R.: Toward Meaningful Simulations of Hadronic Showers. In: M. Albrow & R. Raja (ed.) Hadronic Shower Simulition Workshop. pp. 123–136 (2007)
101. Smirnov, I.B.: Modeling of ionization produced by fast charged particles in gases. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 554, 474–493 (2005). https://doi.org/10.1016/j.nima.2005.08.064

102. Biagi, S.F.: Monte Carlo simulation of electron drift and diffusion in counting gases under the influence of electric and magnetic fields. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 421, 234–240 (1999). https://doi.org/10.1016/S0168-9002(98)01233-9
103. Fedoseyev, A.I., Turowski, M., Alles, M.L. et al.: Accurate numerical models for simulation of radiation events in nano-scale semiconductor devices. Math. Comput. Simul. 79, 1086–1095 (2008). https://doi.org/10.1016/j.matcom.2007.09.013
104. Schrimpf, R.D., Weller, R.A., Mendenhall, M.H. et al.: Physical mechanisms of single-event effects in advanced microelectronics. Nucl. Instruments Methods Phys. Res. Sect. B Beam Interact. with Mater. Atoms. 261, 1133–1136 (2007). https://doi.org/10.1016/j.nimb.2007.04.050
105. Bielajew, A.F., Rogers, D.W.O.: Variance-reduction techniques. In: Jenkins, T.E., Nelson, W.R., Rindi, A., Nalum, A.E., and Rogers, D.W.O. (eds.) Monte Carlo Transport of Electrons and Photons. pp. 407–420. Plenum Press, New York (1990)
106. Abe, F., Akimoto, H., Akopian, A. et al.: Observation of Top Quark Production in p\ifmmode\bar\else\textasciimacron\fi{}p Collisions with the Collider Detector at Fermilab. Phys. Rev. Lett. 74, 2626–2631 (1995). https://doi.org/10.1103/PhysRevLett.74.2626
107. Abachi, S., Abbott, B., Abolins, M. et al.: Observation of the Top Quark. Phys. Rev. Lett. 74, 2632–2637 (1995). https://doi.org/10.1103/PhysRevLett.74.2632
108. Abachi, S., Abolins, M., Acharya, B.S. et al.: The D{Ø} detector. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 338, 185–253 (1994). https://doi.org/10.1016/0168-9002(94)91312-9
109. Collaboration, D., Abazov, V.M., Abbott, B. et al.: Evidence for production of single top quarks. Phys. Rev. D. 78, 12005 (2008). https://doi.org/10.1103/PhysRevD.78.012005
110. Poon, E., Verhaegen, F.: Accuracy of the photon and electron physics in GEANT4 for radiotherapy applications. Med. Phys. 32, 1696–1711 (2005). https://doi.org/10.1118/1.1895796
111. Aharrouche, M., Colas, J., Ciaccio, L. et al.: Energy linearity and resolution of the ATLAS electromagnetic barrel calorimeter in an electron test-beam. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 568, 601–623 (2006). https://doi.org/10.1016/j.nima.2006.07.053
112. Khramov, E., Rusakovich, N., Carli, T. et al.: Study of the Response of the Hadronic Barrel Calorimeter in the ATLAS Combined Test-beam to Pions of Energies from 20 to 350 GeV for Beam Impact Points from 0.2 to 0.65. Geneva (2009)
113. Adragna, P., Alexa, C., Anderson, K. et al.: Measurement of pion and proton response and longitudinal shower profiles up to 20 nuclear interaction lengths with the ATLAS Tile calorimeter. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 615, 158–181 (2010). https://doi.org/10.1016/j.nima.2010.01.037
114. Drozhdin, A.I., Huhtinen, M., Mokhov, N.V: Accelerator related background in the CMS detector at LHC. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 381, 531–544 (1996). https://doi.org/10.1016/S0168-9002(96)00807-8
115. Aaboud, M., Aad, G., Abbott, B. et al.: A measurement of the calorimeter response to single hadrons and determination of the jet energy scale uncertainty using LHC Run-1 pp-collision data with the ATLAS detector. Eur. Phys. J. C. 77, 26 (2017). https://doi.org/10.1140/epjc/s10052-016-4580-0
116. CMS: Determination of jet energy calibration and transverse momentum resolution in CMS. J. Instrum. 6, P11002–P11002 (2011). https://doi.org/10.1088/1748-0221/6/11/P11002
117. Rahmat, R., Kroeger, R., Giammanco, A.: The Fast Simulation of The CMS Experiment. J. Phys. Conf. Ser. 396, 062016 (2012). https://doi.org/10.1088/1742-6596/396/6/062016
118. Grindhammer, G., Rudowicz, M., Peters, S.: The fast simulation of electromagnetic and hadronic showers. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 290, 469–488 (1990). https://doi.org/10.1016/0168-9002(90)90566-O
119. Sekmen, S., Collaboration, for the C.: Recent Developments in CMS Fast Simulation. (2017)

120. Adloff, C., Blaha, J., Blaising, J.-J. et al.: Validation of GEANT4 Monte Carlo models with a highly granular scintillator-steel hadron calorimeter. J. Instrum. 8, P07005 (2013). https://doi.org/10.1088/1748-0221/8/07/P07005
121. Santina, G., Nieminen, P., Evansa, H. et al.: New Geant4 based simulation tools for space radiation shielding and effects analysis. Nucl. Phys. B - Proc. Suppl. 125, 69–74 (2003). https://doi.org/10.1016/S0920-5632(03)90968-6
122. Santin, G., Ivanchenko, V., Evans, H. et al.: GRAS: a general-purpose 3-D Modular Simulation tool for space environment effects analysis. IEEE Trans. Nucl. Sci. 52, 2294–2299 (2005). https://doi.org/10.1109/TNS.2005.860749
123. Gillam, J.E., Rafecas, M.: Monte-Carlo simulations and image reconstruction for novel imaging scenarios in emission tomography. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 809, 76–88 (2016). https://doi.org/10.1016/J.NIMA.2015.09.084
124. Castiglioni, I., Cremonesi, O., Gilardi, M.C. et al.: Scatter correction techniques in 3D PET: a Monte Carlo evaluation. IEEE Trans. Nucl. Sci. 46, 2053–2058 (1999). https://doi.org/10.1109/23.819282
125. Strulab, D., Santin, G., Lazaro, D. et al.: GATE (geant4 application for tomographic emission): a PET/SPECT general-purpose simulation platform. Nucl. Phys. B - Proc. Suppl. 125, 75–79 (2003). https://doi.org/10.1016/S0920-5632(03)90969-8
126. Arce, P., Lagares, J.I., Harkness, L. et al.: GAMOS: An easy and flexible way to use GEANT4. In: 2011 IEEE Nuclear Science Symposium Conference Record. pp. 2230–2237. IEEE (2011)
127. España, S., Herraiz, J.L., Vicente, E. et al.: PeneloPET, a Monte Carlo PET simulation tool based on PENELOPE: features and validation. Phys. Med. Biol. 54, 1723–1742 (2009). https://doi.org/10.1088/0031-9155/54/6/021
128. Barret, O., Carpenter, T.A., Clark, J.C. et al.: Monte Carlo simulation and scatter correction of the GE Advance PET scanner with SimSET and Geant4. Phys. Med. Biol. 50, 4823–4840 (2005). https://doi.org/10.1088/0031-9155/50/20/006
129. Kawrakow, I., Mitev, K., Gerganov, G. et al.: SU-GG-I-109: Using EGSnrc Within GATE to Improve the Efficiency Of positron Emission Tomography Simulations. Med. Phys. 35, 2667–2667 (2008). https://doi.org/10.1118/1.2961507
130. Seco, J.: Monte carlo techniques in radiation therapy. CRC Press (2016)
131. Zacharatou Jarlskog, C., Paganetti, H.: Physics Settings for Using the Geant4 Toolkit in Proton Therapy. IEEE Trans. Nucl. Sci. 55, 1018–1025 (2008). https://doi.org/10.1109/TNS.2008.922816
132. Paganetti, H., Jiang, H., Parodi, K. et al.: Clinical implementation of full Monte Carlo dose calculation in proton beam therapy. Phys. Med. Biol. 53, 4825–4853 (2008). https://doi.org/10.1088/0031-9155/53/17/023
133. Paganetti, H., Jiang, H., Adams, J.A. et al.: Monte Carlo simulations with time-dependent geometries to investigate effects of organ motion with high temporal resolution. Int. J. Radiat. Oncol. Biol. Phys. 60, 942–50 (2004). https://doi.org/10.1016/j.ijrobp.2004.06.024
134. Aso, T., Yamashita, T., Akagi, T. et al.: Validation of PTSIM for clinical usage. In: IEEE Nuclear Science Symposuim & Medical Imaging Conference. pp. 158–160. IEEE (2010)
135. Perl, J., Shin, J., Schümann, J. et al.: TOPAS: An innovative proton Monte Carlo platform for research and clinical applications. Med. Phys. 39, 6818–6837 (2012). https://doi.org/10.1118/1.4758060
136. Böhlen, T.T., Cerutti, F., Dosanjh, M. et al.: Benchmarking nuclear models of FLUKA and GEANT4 for carbon ion therapy. Phys. Med. Biol. 55, 5833–5847 (2010). https://doi.org/10.1088/0031-9155/55/19/014
137. Dedes, G., Pinto, M., Dauvergne, D. et al.: Assessment and improvements of Geant4 hadronic models in the context of prompt-gamma hadrontherapy monitoring. Phys. Med. Biol. 59, 1747–1772 (2014). https://doi.org/10.1088/0031-9155/59/7/1747

138. Mairani, A., Brons, S., Cerutti, F. et al.: The FLUKA Monte Carlo code coupled with the local effect model for biological calculations in carbon ion therapy. Phys. Med. Biol. 55, 4273–4289 (2010). https://doi.org/10.1088/0031-9155/55/15/006
139. Robert, C., Dedes, G., Battistoni, G. et al.: Distributions of secondary particles in proton and carbon-ion therapy: a comparison between GATE/Geant4 and FLUKA Monte Carlo codes. Phys. Med. Biol. 58, 2879–99 (2013). https://doi.org/10.1088/0031-9155/58/9/2879
140. De Napoli, M., Agodi, C., Battistoni, G. et al.: Carbon fragmentation measurements and validation of the Geant4 nuclear reaction models for hadrontherapy. Phys. Med. Biol. 57, 7651–7671 (2012). https://doi.org/10.1088/0031-9155/57/22/7651
141. Battistoni, G., Bauer, J., Boehlen, T.T. et al.: The FLUKA Code: An Accurate Simulation Tool for Particle Therapy. Front. Oncol. 6, 116 (2016). https://doi.org/10.3389/fonc.2016.00116
142. Nikjoo, H., Uehara, S., Emfietzoglou, D. et al.: Track-structure codes in radiation research. Radiat. Meas. 41, 1052–1074 (2006). https://doi.org/10.1016/j.radmeas.2006.02.001
143. El Naqa, I., Pater, P., Seuntjens, J.: Monte Carlo role in radiobiological modelling of radiotherapy outcomes, (2012)
144. Bernal, M.A., Bordage, M.C., Brown, J.M.C. et al.: Track structure modeling in liquid water: A review of the Geant4-DNA very low energy extension of the Geant4 Monte Carlo simulation toolkit. Phys. Medica. 31, 861–874 (2015). https://doi.org/10.1016/J.EJMP.2015.10.087
145. Incerti, S., Douglass, M., Penfold, S. et al.: Review of Geant4-DNA applications for micro and nanoscale simulations. Phys. Medica. 32, 1187–1200 (2016). https://doi.org/10.1016/J.EJMP.2016.09.007
146. Pedoux, S., Cugnon, J.: Extension of the Liège intranuclear cascade model at incident energies between 2 and 12 GeV. Aspects of pion production. Nucl. Phys. A. 866, 16–36 (2011)
147. Dingfelder, M., Ritchie, R.H., Turner, J.E. et al.: Comparisons of calculations with PARTRAC and NOREC: transport of electrons in liquid water. Radiat. Res. 169, 584–594 (2008). https://doi.org/10.1667/RR1099.1
148. Thomson, M.A.: Particle flow calorimetry and the PandoraPFA algorithm. Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip. 611, 25–40 (2009). https://doi.org/10.1016/j.nima.2009.09.009
149. Aduszkiewicz, A., Ali, Y., Andronov, E. et al.: Measurements of $\pi\pm$, K$\pm$, p and p-bar spectra in proton-proton interactions at 20, 31, 40, 80 and 158 GeV/c with the NA61/SHINE spectrometer at the CERN SPS. Eur. Phys. J. C. 77, 671 (2017). https://doi.org/10.1140/epjc/s10052-017-5260-4

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 12
Triggering and High-Level Data Selection

W. H. Smith

12.1 Level-1 Trigger

12.1.1 Introduction

The data taken by a particle physics collider detector consists of events, which are snapshots of the detector data at specific intervals in time. Usually these snapshots are taken at the frequency of the crossing of the colliding beams. For HERA this was 96 ns, for the Tevatron Run II this was 396 ns and for the LHC at design luminosity this is 25 ns. An individual bunch crossing may contain either no, one or many interactions between the particles in the colliding beams. The time during which beam collisions take place during a beam crossing is 1–2 ns. Even if there are multiple collisions in a single crossing the detector elements will make only one recording and the events will be superimposed. Therefore, each bunch crossing is individually evaluated. Not all of the detector data from an individual crossing is available immediately. Some may be stored as charge and need digitization. Other digital detector data may be inaccessible until further detector processing is complete.

The selection of bunch crossings is a highly complex function that involves a series of levels which take increasing amounts of time, process increasing amounts of data, use increasingly complex algorithms and make increasingly more precise determinations to reject increasing numbers of crossings. The first level(s) of the series usually involve(s) specific custom high-speed electronics. The subsequent level(s) involve more general CPU farms that run code similar to that found in the offline reconstruction. Due to this structure, the first level of trigger decision

W. H. Smith (✉)
University of Wisconsin-Madison, Madison, WI, USA
e-mail: wsmith@hep.wisc.edu

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_12

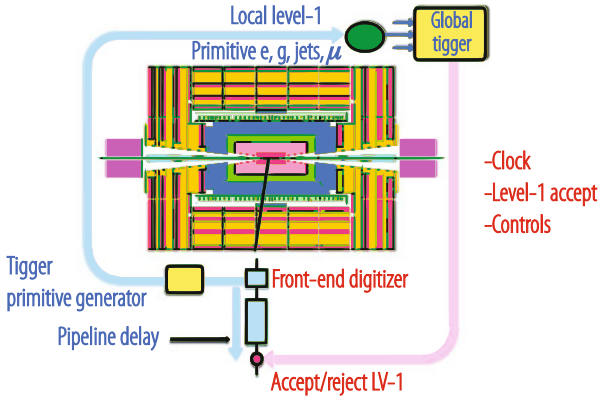


Fig. 12.1 Layout of the elements of the L1T

is based on particle identification (e.g. muon, electron, etc.) from local pattern recognition and energy evaluation. The higher trigger levels start by identifying the particle signature (e.g. Z, W, etc.), calculating kinematics for effective mass and event topology cuts and performing track reconstruction and detector matching (e.g. muon and tracking or calorimeter and tracking). The highest-level triggers perform identification of the physics process detected using event reconstruction and analysis. As shown schematically in Fig. 12.1 the Level-1 trigger[1] (L1T) inspects a subset of the detector information for each bunch crossing and provides the first in a series of decisions to either keep or discard it. The L1T system generally uses coarsely segmented data from calorimeter and muon detectors and in a few cases some rudimentary tracking detector information, while holding all the high-resolution data in pipeline memories in the front-end electronics. During the L1T decision time that is typically a few μsecs, all of the data from all crossings are stored. Usually a good fraction of the L1T time is used in transmission of the L1T data from the detector front ends to a central location where trigger processing is performed and transmission of the L1T decision back to the front ends, leaving a fraction of the L1T decision time available for the trigger processing.

The need to process each new crossing of data requires that the L1T function in a pipelined mode, e.g. be composed of a series of steps each of which processes its input and produces its output result at the crossing frequency. As noted above this can range from 396 ns at the Tevatron to 96 ns at HERA to 25 ns at the LHC. In order to avoid dead time, the trigger electronics must itself be pipelined: every process in the trigger must be repeated at the beam-crossing rate. This has important consequences for the requirements on the structure of the trigger system. The fact that each piece of logic must accept new data at the beam-crossing rate means that no

[1]This is commonly called Level-1 but in such experiments as ALICE and LHCb this corresponds to the Level-0.

piece of individual data processing can take more than this time. This prohibits the use of iterative algorithms, such as jet finding based on finding a seed tower and then adding the surrounding towers to make a jet energy sum. This pipelined structure means that each step in the L1T logic must be completed within the time of the crossing frequency and the results output so that this step in the logic is available to process the data from the next crossing. The L1T logic therefore consists of a number of pipelined steps equal to the processing time multiplied by the crossing frequency.

The tight timing structure of the L1T presents a couple of challenges. Generally, the detector calorimeters have long pulse shapes that exceed the time between beam-crossings. This implies that particles produced in different bunch crossings can produce significant pulse-height in the bunch crossing of interest. Therefore, the detector systems that calculate the input information for the trigger need to correctly identify the energy associated with the correct bunch crossing, usually against a background of additional energy deposits from other bunch crossings. Typically, these systems use peak-finding algorithms and finite input response filters to perform this determination. The gaseous tracking detectors used in the muon systems can also have drift times or pulse widths exceeding the time between bunch crossings. These systems are also required to not only detect the passage of the charged track but also to identify the crossing that produced the track. Often this is resolved by combining and comparing the hits found in adjacent planes of chambers. Another challenge is that the physical extent of large HEP detectors produces times of flight to traverse them that exceed the time between bunch crossings. Therefore, at any particular point in time, the particles from interactions of more than one bunch crossing are present in the detector at different locations. This requires tight timing and synchronization of the detector trigger and readout systems.

The trigger is the start of the physics event selection process. A decision to retain an event for further consideration has to be made at the crossing frequency. This decision is based on the event's suitability for inclusion in one of the various data sets to be used for analysis. The data sets to be taken are determined by the experiment's physics priorities as a whole. Examples of data sets used in LHC experiments include di-lepton and multi-lepton data sets for top and Higgs studies, lepton plus jet data sets for top physics, and inclusive electron data sets for calorimeter calibrations. In addition, other samples are necessary for measuring efficiencies in event selection and studying backgrounds. The trigger has to select these samples in real time along with the main data samples.

The L1T is based on the identification of physics objects such as muons, electrons, photons, jets, taus and missing transverse energy. Each of these objects is typically tested against several p_{T} or E_{T} thresholds. The efficiency of a trigger is determined by dividing the number of events that pass the trigger by the number of actual events that would populate the final physics results plots if all of them passed the trigger. The trigger must have a sufficiently high and understood efficiency at a sufficiently low threshold to ensure a high yield of events in the final physics plots to provide enough statistics and a high enough efficiency for these events so that the correction for this efficiency does not add appreciably to the systematic error of the

measurement. The efficiency of the trigger is evaluated with respect to benchmark physics processes derived from the physics goals of the experiment. The criteria are a sharp turn-on curve of the efficiency at its threshold and an asymptote as close to 100% as possible. The L1T thresholds should be somewhat smaller than the offline physics analysis cuts. The reason for such a requirement is that the efficiency turn-on curves for the L1T will be somewhat softer than can be achieved with a full analysis including the best resolutions and calibration corrections.

Much of the logic in contemporary L1T systems is contained in custom Application Specific Integrated Circuits (ASICs), semi-custom or gate-array ASICs, Field Programmable Gate Arrays (FPGAs), Programmable Logic Devices (PLD), or discrete logic such as Random-Access Memories (RAM) that are used for memory Look-Up Tables (LUT). Given the remarkable progress in FPGA technology, both in speed and number of gates, the technology of many trigger systems has mostly moved towards full implementation in FPGAs.

The key to a good trigger system is flexibility. Not only should all thresholds be programmable, but also as mentioned above, algorithms are either implemented in FPGAs or LUTs. Reprogramming the FPGAs or downloading new LUT contents allows for revisions of the trigger algorithms. The only fairly fixed aspect of the trigger system is which data is brought to which point for processing. However, this is determined by the detector elements, size of showers and curvature of tracks, which are well known and basic features of the detectors and physics signals. There are new technologies being developed that are expected to provide flexibility in data routing, including backplanes and cards that use programmable cross-point switches.

The L1T system sustains a large dataflow. This is either carried on optical fibres, copper cables, or on backplanes within crates. At the LHC, the data carried by these means may be sent in parallel at either 40 MHz, or a higher multiple of this frequency, or converted from parallel to serial and transmitted at a higher rate on a single lines or pair of lines. Serial data transmission has the advantage of transmitting more data per cable wire or backplane pin but the disadvantage of extra latency for the parallel to serial and serial to parallel operations plus the risk of data errors involved with the encoding, high frequency transmission and link synchronization. In many cases this requires the overhead of monitoring and error detection bits. Copper cables in general avoid the necessity for optical drivers with their cost, size and power requirements, but have limited length capability, take up more volume and use more material.

12.1.2 L1T Requirements

The L1T has to be inclusive, local, measurably efficient, and fill the DAQ bandwidth with a high purity stream. The local philosophy of the trigger implies an initial trigger selection of electrons, photons, muons and jets that relies on local information tied directly to their distinctive signatures, rather than on global topologies. For

example, electron showers are small and extremely well defined in the transverse and longitudinal planes. Information from a few Electromagnetic and Hadronic calorimeter towers at the L1T, the corresponding elements of the preshower detector, and a small region of the tracking volume (at higher trigger levels) are sufficient for electron identification. The only global entities are neutrinos (from a global sum of missing E_{T}).

For the trigger to be measurably efficient the tools to measure lepton and jet efficiencies must be built into the trigger architecture from the start. One such tool is overlapping programmable triggers so that multiple triggers with different thresholds and cuts that can run in parallel. A second tool is pre-scaled (e.g. random selection of a fraction) triggers of lower threshold or weaker criteria that run in parallel with the stricter triggers. A third tool is pre-scaling of a particular trigger with one of its cuts removed.

The requirement on the use of DAQ bandwidth implies two conditions. First, each level of the trigger attempts to identify leptons and jets as efficiently as possible, while keeping the output bandwidth within requirements. The selected event sample should include all events that would be found by the full offline reconstruction. Hence, the selection criteria in the trigger must be consistent with those of the offline. Second, since the bandwidth to permanent storage media is limited, events must be selected with care at the final trigger level.

The measurement of trigger efficiency requires the flexibility to have overlapping triggers so that efficiencies can be measured from the data. The overlaps include different thresholds, relaxed individual criteria, prescaled samples with one criterion missing, and overlapping physics signatures. For example, measurement of the inclusive jet spectrum uses several triggers of successively higher thresholds, with the lower thresholds prescaled by factors that allow a reasonable rate to storage. These triggers overlap in jet energy all the way down to minimum bias events so that the full spectrum can be reconstructed accurately. The efficiency and bias of each higher threshold can be measured from the data sets of lower threshold. A requirement for understanding the trigger efficiency is that the data used as input to the L1T system is also transmitted via the DAQ for storage along with the event readout data. In addition, all trigger objects found, whether they were responsible for the L1 trigger or not should also be sent.

The L1T accept rate is limited by the speed of the detector electronics readout and the rate at which the data can be harvested by the data acquisition system. Since it is pipelined and deadtimeless, the L1T renders a decision on every bunch crossing. The maximum L1T accept rate is set by the average time to read information for processing by the Higher Level Triggers (HLT) and the average time for completion of processing steps in the HLT logic.

The high operational speed and pipelined architecture also requires that specific data is brought to specific points in the trigger system for processing and that there cannot be fetching of data based on analysis of other data in an event. The data must flow synchronously across the trigger logic in a deterministic manner in the same way for each crossing. At any moment there are many crossings being processed in sequence in the various stages of the trigger logic. The consequence is that most

of the L1T operations are either simple arithmetic operations or functions using memory lookup tables where an address of data produces a result previously written into the memory.

The L1T requirements evolve with the experiment luminosity, energy and event pile-up (number of p–p collisions per beam crossing). For example for the LHC trigger systems [20], algorithms used by the ATLAS [13] and CMS [14] experiments at the LHC during the period before 2014 (Run-1) were optimized for 7–8 TeV center-of-mass energy, PU up to 40 due to the 20 MHz beam crossing frequency and luminosities up to 7×10^{-33} cm^2 s^{-1}, whereas afterwards (Run-2) these were optimized for a 13 TeV center-of-mass energy and PU above 50 due to the 40 MHz beam crossing frequency and luminosities exceeding 15×10^{-33} cm^2 s^{-1} [15, 16].

12.1.3 Muon Triggers

The design of L1T muon trigger logic depends on the detectors being used to generate the trigger information. These detectors include those with timing resolution and prompt signals that are generally less than the time between bunch crossings such as Resistive Plate Chambers (RPCs) and Thin Gap Chambers (TGCs). They also include special signal handling of detectors with individual signals and resolution greater than the bunch crossing time, such as Cathode Strip Chambers (CSCs) and Drift Tube Chambers (DTs). For these detectors, offset detector planes, front-end logic that processes over the drift time, and combinations of planes provide identification of the bunch crossing associated with the muon passage. Another important feature in muon trigger design is whether the muon chamber measuring stations are placed in a magnetic field in air or embedded in iron. In the former case, the muon momentum resolution is usually sufficient to provide an efficient threshold up to relatively high p_{T}. In the latter case, information from the tracking detectors is needed to provide a sufficiently sharp threshold.

L1T muon algorithms depend on comparison of tracks of hits with predefined geometrical patterns such as roads. For example, the ATLAS muon trigger employs RPCs and TGCs in an air-core magnetic field and the trigger algorithm uses Coincidence Windows that start with a hit in a central "pivot plane" and searches for time-correlated hits within an η–ϕ window in a "confirm plane" [1]. Different "confirm planes" are used for low and high p_{T} muons, as is shown in Fig. 12.2. The RPC barrel algorithm extrapolates hits in the middle RPC 2 station to a point and coincidence window in the innermost RPC 1 station along a straight line to the nominal interaction point. The size of this coincidence window depends on the muon's bend in the magnetic field. A low-p_{T} candidate is found if there is one hit in this window and hits in both views and planes of either RPC 1 or RPC 2. If there is also a hit in RPC 3, then a high-p_{T} candidate has been found. For Run 2 ATLAS commissioned a fourth layer of barrel RPCs that improved the acceptance and added new trigger logic to the end-cap requiring additional coincidences with the TGC's

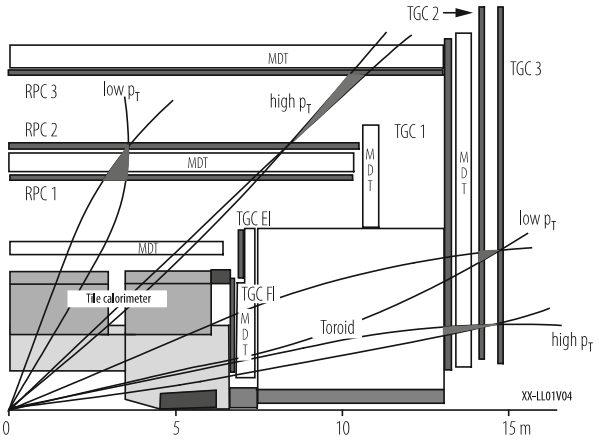


Fig. 12.2 ATLAS muon trigger algorithms

or the Tile hadronic calorimeter to reject particles not originating at the interaction point [15].

The CMS Detector uses Drift Tubes (DT), Cathode Strip Chambers (CSC) and overlapping Resistive Plate Chambers (RPC) for muon triggering in iron. The RPC readout strips are connected to pattern logic, which is projective in η and ϕ and connected to segment processors that find the tracks and calculate the p_T. As shown in Fig. 12.3 the CSC logic forms Local Charged Tracks (LCT) from the charge distributions in the CSC planes, which are combined with the Anode wire information for bunch crossing identification and assignment of p_T and "quality", which is an indicator of the number of planes hit. The CSC Track Finder combines the LCTs into full muon tracks and assigns p_T values to them. As is also shown in Fig. 12.3 the DTs are equipped with Bunch and Track Identifier (BTI) electronics that finds track segments from coincidences of aligned hits in four layers of one drift tube superlayer. The DT Track Finder combines the segments from different stations into full muon tracks and assigns p_T values to them. In Run 1, the Global Muon Trigger sorted and then correlated the RPC, DT and CSC muon tracks. In Run 2, the RPC, DT and CSC information were combined earlier, in the track-finding stage [12].

The LHCb Level-0 muon trigger searches for candidates in the quadrants of five stations of Multi-Wire Proportional Chambers separated by iron and sends the two highest p_T candidates from each quadrant to the Level-0 Decision Unit (L0DU) [2]. The ALICE dimuon trigger system is based on two stations of 18 RPCs each read out on both sides of the gas gap by X–Y orthogonal strips with high resolution front-end electronics which feed local trigger electronics modules that find tracks in 3 out of the 4 detector planes in both X and Y [3]. The track is found and the magnetic deviation is calculated to enable a cut on a p_T threshold using memory Look-Up Tables (LUTs). Two unlike-sign muons are then required in the L1T.

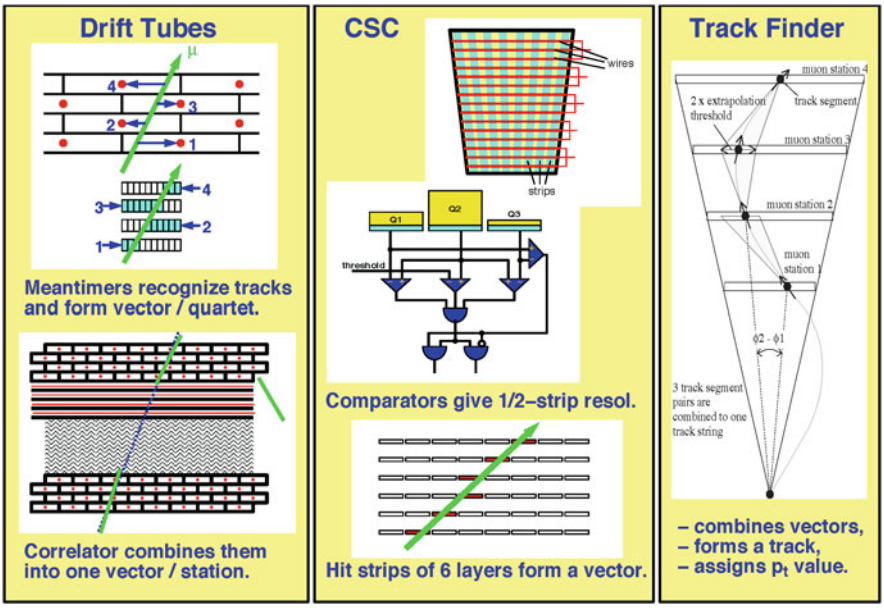


Fig. 12.3 CMS muon chamber trigger algorithms

12.1.4 Calorimeter Electron and Photon Triggers

The calorimeter trigger begins with trigger tower energy sums formed by the detector electromagnetic calorimeter (ECAL) hadronic calorimeter (HCAL) and forward calorimeter. Experiments vary on whether these sums are performed by analog methods before digitization or by digital summation after an initial ADC.

For the ATLAS experiment, the calorimeter trigger begins with a Preprocessor (PPr) which sums analog pulses into 0.1×0.1 ($\eta \times \phi$) trigger towers, assigns their bunch crossing and adjusts for calibration. The Cluster Processor then identifies and counts electron/photon and tau candidates based on the energies and patterns of energy isolation found in overlapping windows of 4×4 ECAL and HCAL trigger towers as shown in Fig. 12.4. For Run 2, the PPr was upgraded to provide improved Finite Input Response (FIR) filtering and dynamic bunch by bunch pedestal correction [15]. New cluster merging modules (CMX) were added that transmitted the location and energy of trigger objects, rather than the threshold multiplicities used in Run 1.

The CMS Calorimeter trigger algorithm for electron and photon candidates uses a 3×3 trigger tower sliding window centered on all ECAL/HCAL trigger towers. A diagram of this electromagnetic algorithm is shown in Fig. 12.5. Two types of electromagnetic objects are defined. The non-isolated electron/photon identification is based on a large energy deposit in one or two adjacent ECAL 5-cell ϕ strips in the trigger tower, the lateral shower profile in the central tower comparing maximum E_T

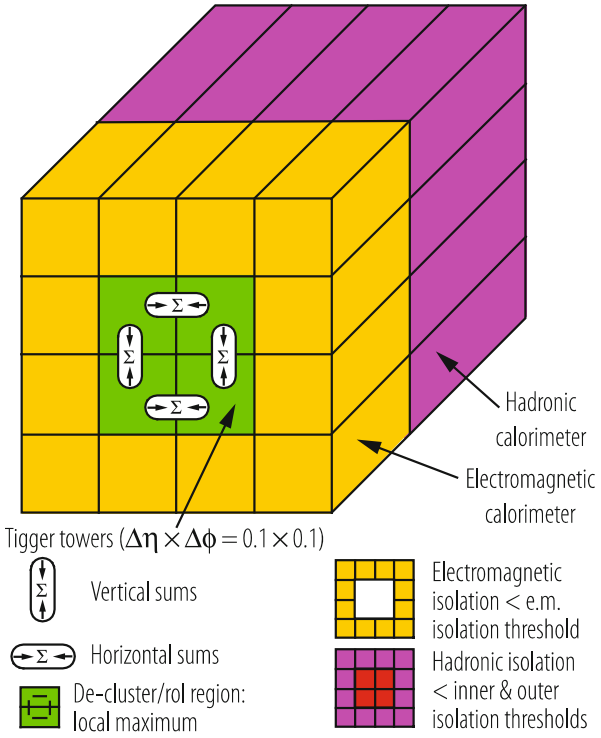


Fig. 12.4 ATLAS calorimeter electron/photon trigger algorithm

of each of four pairs of strips of 5 cells to the total tower level E_T of all 25 crystals (this "Fine Grain" veto uses a strip due to electron bending in the magnetic field), and the longitudinal shower profile defined by the ratio of E_T deposits in the HCAL and ECAL portions of the calorimeter (H/E veto). The isolated electron/photon has two additional requirements: the ECAL E_T deposited in one of the five trigger towers surrounding the central tower is below a programmable E_T threshold and the eight trigger towers surrounding the central tower in the 3×3 region have passed the Fine Grain and H/E vetoes. For Run-2, the CMS Calorimeter Trigger hardware was upgraded so that more complex algorithms could be deployed [21]. The e/γ and τ candidates started with a local maximum around which the trigger towers were dynamically clustered.

The LHCb Level-0 calorimeter trigger system combines the E_T measurement in clusters of 2×2 cells in the electromagnetic (ECAL) and hadronic calorimeters (HCAL), as well as information from the Scintillator Pad Detector (SPD) and a Preshower (Prs) to indicate the charged and electromagnetic nature of the clusters. The calorimeter trigger system sends the highest E_T hadron, electron, photon and π^0 candidates and the total HCAL E_T and SPD multiplicity to the Level 0 Decision Unit (L0DU).

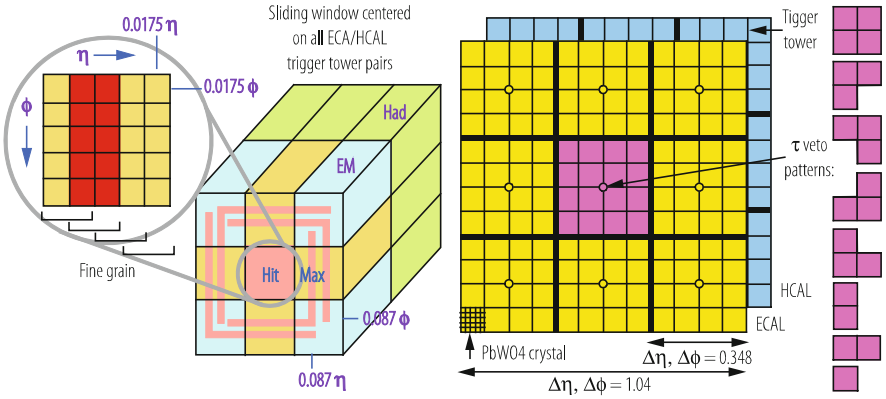


Fig. 12.5 CMS calorimeter electron/photon and jet/tau trigger algorithms

12.1.5 Calorimeter Jet and Missing Energy Triggers

The Level-1 Calorimeter Jet trigger needs to approximate the offline and higher-level trigger iterative jet finding in cones around seed towers with rectangular sliding windows of trigger towers. As shown in Fig. 12.5, the CMS jet trigger algorithms are based on sums of 3 × 3 calorimeter regions. This corresponds to 12 × 12 trigger towers in the barrel and endcap where a region corresponds to 4 × 4 trigger towers. The algorithm uses a 3 × 3 sliding window technique that uses the complete (η, ϕ) coverage of the CMS calorimeter. The E_T of the central region is required to be higher than that of the eight neighbours. The central jet or τ-tagged jet is defined by the 12 × 12 trigger tower E_T sum. In the case of τ-tagged jets, none of the nine 4 × 4 regions are allowed to have energy deposited outside the patterns of ECAL or HCAL towers (i.e. above a programmable threshold). For Run-2 the upgraded CMS calorimeter trigger formed Jet candidates by grouping the trigger towers around a local maximum in a 9 × 9 tower region in η × φ with a PU subtraction estimated using four surrounding 3 × 9 tower regions. The ATLAS Jet and Energy L1T algorithm is based on a sliding window of 4 × 4 sums of trigger towers. It operates on a 4 × 8 matrix of core towers as shown in Fig. 12.6. In order to perform its calculations, it also needs the energy deposited in the "environment" of 7 × 11 towers. The execution of this algorithm depends on the duplication and distribution of energies in order to supply the needed information to perform these sums.

12.1.6 Tracking Information in Level-1 Triggers

Tracking information is very effective in reducing backgrounds to level-1 electron triggers from π^0s. It improves tau triggers by identifying isolated tracks and it

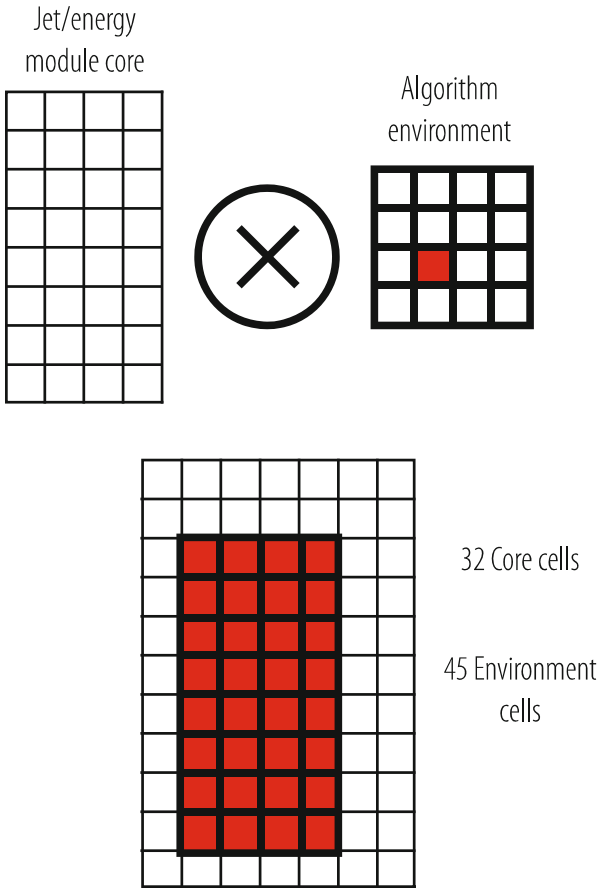


Fig. 12.6 Organization of the ATLAS jet trigger system

refines the muon trigger with a sharper momentum threshold that is not affected by the backgrounds in the muon chambers. It also can be used to identify heavy flavour candidates. Both Tevatron experiments CDF and DØ employed level-1 tracking triggers. CDF used signals from the Central Outer Tracker (COT) open-cell drift chamber in the eXtremely Fast Tracker (XFT) to perform charged track reconstruction in the r–ϕ plane for the L1T [4]. Track segments were found by comparing hit patterns in a COT superlayer to a list of valid patterns or "masks". These masks contained specific patterns of prompt and delayed hits on the 12 wire layers of an axial COT superlayer. Tracks were found by comparing track segment patterns in all four layers to a list of valid segment patterns or "roads". The XFT had an efficiency >90% for tracks with $p_{\mathrm{T}} > 1.5$ GeV/c, transverse momentum resolution of $\delta p_{\mathrm{T}}/p_{\mathrm{T}} = 0.002\, p_{\mathrm{T}}$ and pointing resolution of $\delta\phi = 0.002$ radians with respect to the beam line [5]. The XFT reported the highest p_{T} track in each of 288 azimuthal

segments (1.25° each) to the XFT "Linker system" modules which cover 15° each and are matched to the segmentation of the trigger signals from the muon and calorimeter systems. The results from the linker system were passed to the Track Extrapolation System (XTRP), which sent one or more bits in 2.5° segmentation to the muon trigger systems set according to the calculated p_T, ϕ and multiple scattering. The XTRP also sent a set of 4 bits (for four momentum thresholds) for each 15° calorimeter wedge to the Level-1 calorimeter trigger. Finally the XTRP created a Level-2 tracking trigger based on the number of tracks and their p_T and ϕ information.

The DØ experiment Central Tracking Trigger (CTT) used information from the Central Fiber Tracker (CFT) and the Central Preshower System (CPS). Hit information from each of the 80 axial sectors of the CFT/CPS detectors was fed through boards programmed with 16,000 Boolean equations that identified patterns of hits likely to be produced by a charged particle. A list of tracks in four momentum ranges between 1.5 and 10 GeV/c was then sent to the L1 muon trigger system [6]. The DØ L1 CTT also identified the number of tracks in each event for each of these four momentum ranges, whether a coincident CPS hit had been found, and whether the track was isolated. This information was also used in the DØ L1T decision. The DØ CTT had an efficiency of $97.3 \pm 0.1\%$ for tracks with $p_T > 10$ GeV/c [4].

Although both ATLAS and CMS are planning the use of Tracking information at Level-1 in their designs for the High Luminosity LHC (HL-LHC) project [22], this information was not included in Run-1 or Run-2.

12.1.7 Global Triggers

An experiment Global Trigger accepts muon, calorimeter and tracking (if available) trigger information, synchronizes matching sub-system data arriving at different times and communicates the Level-1 decision to the timing, trigger and control system for distribution to the sub-systems to initiate the readout. The global trigger decision is made using logical combinations of the input trigger data. Besides handling physics triggers, the Global Trigger provides for test and calibration runs, not necessarily in phase with the machine, and for prescaled triggers, as this is an essential requirement for checking trigger efficiencies and recording samples of large cross section data.

The ATLAS Level-1 Global trigger is called the Central Trigger Processor (CTP). It combines information on the multiplicities of calorimeter and muon trigger objects which have sufficiently high momentum. These are electrons/photons, taus, jets, and muons. These are also the "seeds" for the Level-2 trigger that are sent to the Region of Interest Builder (RoIB). In addition, threshold information on the global transverse energy and missing energy sums is also used in the Level-1 decision. In Run-1, the CTP discriminated the delivered multiplicities of the trigger objects against multiplicity conditions and then combined these conditions to form more complex triggers when multiple object triggers are needed. In Run-2, the ATLAS

L1 Global trigger added a topological trigger (L1Topo) to allow geometrical or kinematic association between trigger objects received from the L1 Calorimeter or Muon Triggers [23].

The CMS L1 Global Trigger sorts ranked trigger objects, rather than histogramming objects over a fixed threshold. This allows all trigger criteria to be applied and varied at the Global Trigger level rather than earlier in the trigger processing. All trigger objects are accompanied by their coordinates in (η, ϕ) space. This allows the Global Trigger to vary thresholds based on the location of the trigger objects. It also allows the Global Trigger to require trigger objects to be close or opposite from each other. In addition, the presence of the trigger object coordinate data in the trigger data, which is read out first by the DAQ after a Level-1 trigger accept (L1A), permits a quick determination of the regions of interest where the more detailed HLT analyses should focus. The Global L1 Trigger transmits a decision to either accept (L1A) or reject each bunch crossing. This decision is transmitted through the Trigger Throttle System (TTS) to the Timing Trigger and Control system (TTC). The TTS allows the reduction by prescaling or blocking L1A signals in case the detector readout or DAQ buffers are at risk of overflow. For Run-1, the Global L1 Trigger allowed up to 128 algorithms to contribute to the overall trigger decision. For Run-2, this was upgraded to a modular design capable of up to about 500 algorithms that was typically running about 300 [16].

12.2 Higher-Level Selection

12.2.1 Introduction

The design of the Higher-Level Selection of events after Level-1 takes place in a number of "trigger levels". Generally, collider experiments use at least two additional trigger levels, referred to as the Level-2 and Level-3 trigger. Some experiments have a Level-4 trigger. The higher the number the more general purpose (or commercial) the implementation, with the Level-3 and Level-4 triggers being composed of farms of standard commodity computers. The physical implementation of the Level-2 trigger varies substantially between experiments from inclusion in the Level-3 farm of processors to an independent farm of processors to customized dedicated processing hardware. The Level-2 trigger has to operate at the output rate of the Level-1 trigger, generally with a subset of the higher resolution and full-granularity available to the full reconstruction code available at Level-3 and higher. Typically, the Level-1 output rate ranges between 1 and 100 kHz depending on the experiment. The Level-2 trigger is generally limited in execution time so that the full event data cannot be unpacked and processed. Instead, the higher resolution and full-granularity data is unpacked in "regions of interest" determined by the Level-1 trigger data.

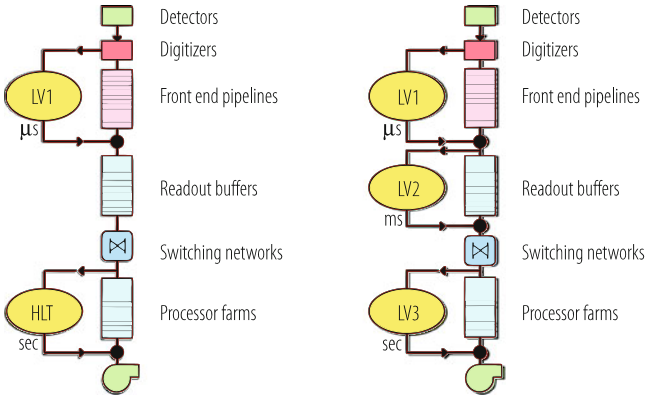


Fig. 12.7 Common architectures for collider detector trigger and data acquisition systems. Left: two physical levels. Right: three physical levels

The architectures of Level-2 trigger systems vary depending on the rejection factor required, the information provided as input, the interconnections with the front-end electronics, Level-1 and Level-2. Examples of two types of architecture presently employed by general-purpose collider detectors are shown in Fig. 12.7. Including Level-1, experiments such as H1, ZEUS, CDF, DØ and ATLAS have three physical levels of processing [18]. For Run-2, the ATLAS Higher Level Trigger Layers (HLT) were combined [15]. CMS has two layers of physical processing [19]. LHCb has three levels of processing, but the first level (Level-0) output trigger rate is 1.1 MHz, an order of magnitude higher than other collider experiments [7, 17]. The subsequent levels, HLT1 and HLT2, are software-based, running on the Event Filter Farm. In Run-2, HLT1 and HLT2 became two independent asynchronous processes on the same node and HLT2 was able to run a full reconstruction on real-time aligned and calibrated data [17].

There are more substantial differences in trigger architecture for experiments such as ALICE that is designed to study heavy ion collisions with a bunch spacing of 125 ns at a lower luminosity than the LHC experiments ATLAS and CMS. However, each Pb–Pb collision produces much higher multiplicities of secondary particles than a p–p collision, resulting in a much higher event size. Since the detectors in ALICE have different readout times, there are three parallel trigger systems, allowing readout from the faster detectors while slower detectors are occupied with reading out the data from earlier events [8]. The first decision is made 1.2 μs after the event (Level-0), the Level-1 decision comes after 6.5 μs, and the Level-2 trigger is issued after 88 μs. The Level-1 and Level-2 decisions can veto trigger signals from Level-0. The ALICE Central Trigger Processor also checks the events for pile-up from events in a programmable time interval before and after the interaction at all three levels. For Run-2, an earlier L0 trigger decision time of 525 ns provides a pre-trigger for the TRD [24].

The algorithms deployed in the HLT are dynamic, reflecting continuing improvements in the offline reconstruction that represent the functions that the HLT is attempting to approach within the constraints of processing time. The descriptions below of algorithms in the LHC experiments represent a snapshot at the time of Run-1 processing. These considerably evolved during Run-1 into Run-2, although the general techniques shown continue to be applied.

12.2.2 Tracking in Higher Level Triggers

The principal new information in the higher level triggers is tracking information. Either it is introduced for the first time in the event selection process or it is greatly refined over rudimentary tracking used in the Level-1 trigger. There are two major sources of tracking information. A pixel detector provides the most inner tracking and some vertex information. Outside of the pixel detector, silicon strip and then in some cases drift chambers, fibers, or straw tube detectors provide additional information at larger radius. For example, ATLAS [9] uses space points found in the pixels and silicon central tracker (SCT) to find the z-vertex location, fit tracks into the Transition Radiation Tracker (TRT) and measure the ϕ and p_{T} of the track above a p_{T} of 0.5 GeV/c. In the latter part of Run-2, ATLAS commissioned the Fast TracKer (FTK), a dedicated Associative-Memory hardware processor which delivers tracks with $p_{\mathrm{T}} > 1$ GeV/c for every L1A to the HLT within 100 μs [25]. In CMS, two types of tracking are employed. Charged particle tracks are first quickly reconstructed using pixel hits and then more laboriously but more accurately reconstructed with additional hits from the silicon strip tracker. Generally, tracking is "seeded" by the confirmed Higher Level Trigger objects, which themselves are "seeded" by Level-1 trigger objects.

12.2.3 Selection of Muons

The first algorithms executed in Level-2 on Level-1 selected muons are refinements of the reconstruction of the tracks in the muon chambers. In the case of ATLAS, where only the RPC (Barrel) and TGT (Forward) chambers provide information for the L1T, the precision hit information from the Monitored Drift Tubes (MDTs) is added to the RPC and TGC determined candidates. This provides good track reconstruction in the muon spectrometer. Since the ATLAS Muon Chambers are mostly in air, there is little multiple coulomb scattering. The found tracks are extrapolated for combination with tracks found in the Inner Detector. Matching between muon tracks measured independently in the muon system and those in the Inner Detector selects prompt muons and rejects fake and secondary muons. The isolated muon triggers also use information from the calorimeter towers surrounding the found muon track.

In CMS, all of the muon chamber systems participate in the L1T. The L1T muon candidates are used to seed the reconstruction of tracks in the muon chambers in the Level-2 algorithm. First, an initial pattern recognition is performed on muon segments along the trajectory, then a second more precise fit using all hits on these segments is used to determine the muon parameters. Since the CMS chambers are surrounded by steel, the propagation of track parameters to adjacent muon stations must take into account material effects such as multiple Coulomb scattering, and energy losses due to ionization and bremsstrahlung in the muon chambers and the iron. To avoid excessive processing times, these are estimated from fast parameterizations. Muons passing this first reconstruction are then input to the Level-3 reconstruction that uses hits in the silicon tracker within a rectangular $\eta \times \phi$ region. Pairs or triplets of hits in the innermost layers of the tracker form trajectory seeds that are required to be compatible with the $\eta \times \phi$ region and the primary vertex constraints. These are then grown into tracks of about seven hits and optionally combined with the reconstructed hits from the Level-2 algorithm. In Level-2, the isolation variable is calculated from the weighted sums of energies deposited in the ECAL and in the HCAL in the region around the muon track. For the Level-3 isolation variable, only charged-particle tracks near the vertex of the candidate muon are selected for inclusion. This excludes tracks from pile-up of contributions from other pp collisions (which occur at another vertex location), making this isolation less sensitive to pile-up than calorimetric isolation.

12.2.4 Selection of Electrons and Photons

The first algorithms executed in Level-2 on Level-1 selected electrons and photons are refinements of the clustering algorithms. For example, in ATLAS [9], the energy deposited in windows of the electromagnetic LAr calorimeter cells and the energy-weighted position information, as well as the leakage energy into the hadronic calorimeter are calculated. CMS [10] also reconstructs energy in clusters of electromagnetic calorimeter cells corresponding to the Level-1 calorimeter triggers, adding a margin around the trigger region to ensure complete collection of energy. These clusters are then formed into "Super Clusters" which are groups of clusters along a road in the ϕ direction, chosen due to bending in the magnetic field. These clusters are then required to be isolated in the electromagnetic calorimeter. The hadronic calorimeter energies are then reconstructed and the energies in the hadronic tower behind the cluster and the adjacent towers are required to be small with respect to the electromagnetic cluster energy.

The second tier of algorithms performed on electrons and photons confirmed by the first algorithms are tracking algorithms. The first or more local steps of these are generally called Level 2.5 algorithms. This involves establishing track isolation around the electromagnetic cluster and for electron triggers, associating the electromagnetic cluster with a track. For CMS electron triggers, the energy and position of the Super Cluster is used to search for hits in the pixel detector. These

hits are reconstructed and the track p_T is checked for consistency with the Super Cluster energy. For both electron and photon triggers, tracks are seeded from pairs of hits in the pixel layers in a rectangular $\eta \times \phi$ region around the direction of the reconstructed electron or photon, where these seeds are required to be consistent with the nominal vertex spread (photons) or closest approach of the electron path to the beam line (electrons). Then for electrons a threshold is applied to the p_T sum of the tracks within a cone around the electron direction and on the number of tracks for the photon. In ATLAS [11] the electromagnetic cluster is identified as an electron by association with a track in the Inner Detector, which is found by independent searches in the SCT/Pixel and TRT detectors in the region identified by the L1T RoI. For electron candidates, matching in both position and momentum between the track and cluster is required.

12.2.5 Selection of Jets and Missing Energy

The primary processing of the jet candidates at Level-2 begins with the L1T jet candidates, which are used as seeds for the Level-2 jets. The first step is to recalculate the jet energy for these candidates using the full granularity and calorimeter energy resolution information, which is not available to the Level-1 jet energy calculation. In ATLAS, the Level-2 jet finding searches in the RoIs produced by the Level-1 calorimeter logic. In CMS, jets are reconstructed using an iterative cone algorithm with cone size $R = \sqrt{\Delta\eta^2 + \Delta\phi^2} = 0.5$ that sums over all projected electromagnetic and hadronic calorimeter cells with energy greater than a threshold set above the level of noise (0.5 GeV). In addition, to be declared a jet, at least one seed tower must have $E_T > 1$ GeV. After summation, the jet energy is adjusted by an η-dependent correction for the calorimeter response.

Missing energy is calculated by summing all towers with E_T above a noise threshold. For CMS, this threshold is 0.5 GeV. No energy corrections are applied to Missing E_T. Since Missing E_T is susceptible to noise because it is summing over many channels, an alternative is often considered. This is Missing H_T, which is Missing E_T calculated by summing over the jets in the event rather than the calorimeter cells. Since there are fewer cells involved in the computation of missing H_T, there is less noise included in this sum.

It is typical to ask for two or more jets in the HLT algorithms. It is also common to combine two or more jets with missing E_T or H_T. Also, topological constraints are often employed such as requiring forward jets or acoplanarity between multiple jets or jets and Missing E_T.

12.2.6 Selection of Hadronic Tau Decays

The Level-2 processing of tau jets relies only on calorimeter information. In ATLAS, the tau finding uses the same algorithms used for electron and photon candidates, but retuned for taus. The inputs are the Level-1 RoIs. A cluster summed over the full resolution data for the electromagnetic and hadronic cells is required to have $E_{\mathrm{T}} > 20$ GeV with at least 10 GeV required individually in the electromagnetic and hadronic cells. The position of the candidate cluster is required to be consistent with the Level-1 tau-jet candidate. Then shower shape variables are used to discriminate tau jets from regular jets. An example of one such variable is R_{37}, defined as the ratio of E_{T} contained in a 3×7 cell cluster to the E_{T} contained in a 7×7 cell cluster centred on the same seed cell calculated for the second electromagnetic layer of the LAr calorimeter. In CMS, the Level-1 tau jets are used as seeds for the Level-2 tau-jet reconstruction that employs an iterative cone algorithm with a radius of $R = 0.5$. Level-2 tau candidates are then these jets which have $E_{\mathrm{T}} > 15$ GeV and are tagged as isolated if the sum of electromagnetic calorimeter deposits in an annulus $0.13 < R < 0.4$ around the jet direction, $E_{\mathrm{T}} < 5$ GeV.

The subsequent processing of tau candidates involves tracking. ATLAS requires a track formed from the pixel and SCT detector space points in the RoI to be within $\Delta R < 0.3$ of the Level-2 tau candidate cluster direction. At Level-3 a requirement is made that the number of tracks within $\Delta R < 0.3$ be either one or three. Additional detailed jet shape requirements also refine the identification. In CMS, at Level 2.5 (the higher level trigger processing following the initial Level-2 processing that uses calorimeter and muon information alone), tau selection is based on tracks with a $p_{\mathrm{T}} > 5$ GeV/c that are reconstructed from seeds from the pixel hits found in a small rectangle ($\Delta\eta = \Delta\phi = 0.1$) around the tau-candidate direction. At Level 3 the rectangle is expanded to 0.5 and the p_{T} cut is reduced to 1 GeV/c. To save CPU time, these tracks are terminated when seven hits in the silicon strip tracker are acquired since the resolution with seven hits is close to final. Reconstructed tracks are associated with the tau-jet candidate if they are within a radius $R < 0.5$ and originate from the primary vertex as determined by the pixel tracks. Tracks within a radius $R < 0.1$ of the tau-jet candidate direction are classed as tau tracks. The leading tau track must have $p_{\mathrm{T}} > 3$ GeV and there must be no reconstructed tracks within an annulus $0.07 < R < 0.3$ around this track.

12.2.7 Selection of b-Jets

The b-jet selection is based on track reconstruction to tag displaced vertices associated with the jet. In ATLAS at Level-2, b-tagging uses reconstructed tracks from the silicon tracker within the Level-1 jet RoI. For each of these tracks the significance of the transverse impact parameter is computed and its error is

parameterized as a function of p_T. A b-jet discriminator is constructed using the likelihood ratio method to determine for each track in the jet the ratio of probability densities for the track to come from a b-jet or a u-jet. In CMS, Level-2 starts with events with 1, 2, 3 or 4 jets passing various thresholds or a high total E_T for the whole event. At Level 2.5 tracks are reconstructed using only pixel hits (at least three required), which are used to reconstruct the primary vertex. The b-tag algorithm runs on the four highest E_T jets with $E_T > 35$ GeV and uses the pixel tracks and primary vertex to tag jets as b-jets if they have at least two tracks with a signed 3D impact parameter with large significance. Events pass Level 2.5 if they have at least one b-tagged jet. At Level 3, tracks of up to eight hits are reconstructed in a cone of size $\Delta R = 0.25$ around the b-tagged jets. The level-3 filter selects events where there is at least one jet having at least two tracks with large impact parameter significance.

12.3 Outlook

Trigger and DAQ requirements will further evolve in the next decade with large increases in luminosity and the associated pile-up. ALICE will continuously read out the majority of its detectors with different latencies, busy times and technologies, differently optimized for pp, pA and AA running scenarios [26]. Triggered readout will be used by some detectors and for commissioning and some calibration runs. LHCb will run trigger-free at 30 MHz, reading every bunch crossing with inelastic collisions [27].

A major upgrade to the LHC, the HL-LHC [28], is planned to start in the middle of this decade and deliver a luminosity of 5–7 $\times$ 10^{34} cm^{-2} s^{-1} at the LHC design centre of mass energy of 14 TeV, which corresponds to a pile-up of 140–200 at 25 ns bunch spacing. Present link technologies operable in the radiation and magnetic field environments of their inner detectors do not allow ATLAS and CMS to adopt a "triggerless" architecture with an acceptable detector power and material budget for their tracking detectors. Therefore, at the HL-LHC, both ATLAS and CMS will retain architectures with Level 1 triggers.

In order to maintain Run-2 physics sensitivity at the HL-LHC, ATLAS and CMS will add L1 tracking triggers for identification of tracks associated with calorimeter and muon trigger objects and will also feature a significant increase of L1 rate, L1 latency and HLT output rate. Additionally, ATLAS and CMS are also studying the use of fast timing information in the L1T. The ATLAS experiment will divide its L1T into two stages [29]. A L0 trigger with a rate of 1 MHz and latency of 6 μs will use calorimeter and muon trigger information to produce seeds used with tracking and more fine-grained calorimeter and muon trigger information in the L1 trigger with an output rate of 400 kHz and latency of 30 μs. This is processed by the HLT with an output storage rate of 5–10 kHz. The CMS L1T latency will increase to 12.5 μs with an output range of 500–750 kHz for pileup ranging between 140 and 200 [30]. It will use an un-seeded L1 Track trigger along with finer granularity

calorimeter and muon triggers. The CMS HLT output rate to storage will range between 5 and 7.5 kHz for pileup ranging between 140 and 200.

The hardware implementations of the HL-LHC ATLAS and CMS L1T will use high-bandwidth serial I/O links for data communication and large, modern field-programmable gate arrays (FPGAs) for sophisticated and fast algorithms. The development and synthesis of FPGA firmware incorporating these algorithms is significantly enhanced in reliability, accessibility and performance with Higher Level Synthesis (HLS) tools [31]. The latest developments and expectations for future FPGAs not only include significant increases in the number of logic gates available and high-speed serial links, but also increases in the number of high-bandwidth serial links per device, more sophisticated and fast DSPs, embedded Linux, and integration with high speed networking. Fast Tracking Trigger devices such as the ATLAS FTK [25] use Associative Memories. The hardware framework will be designed following standards deployed in industry, such as the Advanced Telecommunications Architecture (ATCA) for backplanes, which offers substantial backplane bandwidth and flexibility and provides for users to extend the backplane connectivity using the spare I/O available on each card. Further interconnectivity technology developments such as optical backplanes and wireless data transmission may provide additional opportunities.

The increase in L1 output rate from 100 kHz to possibly as high as 1 MHz requires higher bandwidth into the DAQ system and more CPU power in the HLT. The addition of a tracking trigger and more sophisticated algorithms at L1 increases the purity of the sample of events passing the L1 trigger, but requires a higher sophistication and complexity of algorithms used at the HLT. This implies a greater CPU power than scaling with the L1 output rate but is somewhat mitigated by the availability of the L1 Tracking Trigger primitives in the data immediately accessible by the HLT. Without a L1 tracking trigger, the opportunity to access most of the tracker information at the first levels of the HLT is limited by the CPU time to unpack and reconstruct the tracking data. This is significantly improved in the ATLAS FTK that provides quick access to tracking information in the HLT. For the HL-LHC, the addition of the L1 tracking trigger means that the results from the L1T track reconstruction can be immediately used without the overhead of tracking data unpacking and reconstruction.

The evolution of the computing market towards different computing platforms and co-processors offers an opportunity to achieve substantial gains in HLT processing power at the price of adapting code to the new hardware. Examples include Graphical Processor Units (GPUs), such as the NVIDIA Tesla and GeForce (used by ALICE [32]), ARM processors, FPGAs (e.g. the Xeon/FPGA used by LHCb [33]) and the Intel Xeon Phi coprocessor. Additional HLT processing power may result from improved code such as machine learning algorithms for track reconstruction [34].

References

1. M. J. Woudstra and the ATLAS Collaboration, Performance of the ATLAS muon trigger in pp collisions at $\sqrt{s} = 8$ TeV, 2014 J. Phys.: Conf. Ser. 513 012040.
2. E. Aslanides et al., (LHCb Collaboration), The Level-0 muon trigger for the LHCb experiment, Nucl. Instr. and Meth. A579 (2007) 989–1004.
3. B. Forestier, Nucl. Instrum. Meth. A 533 (2004) 22–26.
4. R. Downing et al., Nucl. Instrum. Meth. A 570 (2007) 36–50.
5. E.J. Thompson et al., *Online Track Processor for the CDF Upgrade*. IEEE Trans. Nucl. Sci. 49(3) (2002).
6. Y. Maravin et al., *First Results from the Central Tracking Trigger of the DØ Experiment.* Proc. 2003 IEEE Nuclear Science Symposium, Portland, OR.
7. R. Aaij et al., (LHCb collaboration), *The LHCb Trigger and its Performance in 2011*, 2013 JINST 8 P04022.
8. M. Krivda et al. (ALICE Collaboration), *The ALICE trigger system performance for p-p and Pb-Pb collisions* 2012 JINST 7 C0105.
9. I. A. Christidi and the ATLAS Collaboration, *The tracking performance of the ATLAS High Level Trigger in pp collisions at the LHC*, 2011 J. Phys.: Conf. Ser. 331, 032006.
10. The CMS Collaboration, *The Trigger and Data Acquisition project*, Vol. II: *Data Acquisition and the Higher Level Trigger*. CERN/LHCC 2007-021.
11. Will Buttinger (ATLAS collaboration) *The ATLAS Level-1 Trigger System*, 2012 J. Phys.: Conf. Ser. 396 012010.
12. D. Acosta et al. (CMS Collaboration), *CMS Trigger Improvements Towards Run II*, Nuclear and Particle Physics Proceedings 273–275 (2016) 1008–1013.
13. The ATLAS Collaboration, *Performance of the ATLAS Trigger System in 2010*, Eur. Phys. J.C 72, 1849 (2012).
14. V. Khachatryan et al., (CMS Collaboration) *The CMS Trigger*, 2017 JINST 12 P01020.
15. The ATLAS Collaboration, *Performance of the ATLAS Trigger system in 2015*, Eur. Phys. J. C (2017) 77: 317.
16. L. Cadamuro et al. (CMS Collaboration), The CMS Level-1 trigger system for LHC Run II, 2017 JINST 12 C03021.
17. B. Sciascia (LHCb Collaboration), *LHCb Run 2 trigger performance*, PoS BEAUTY2016 (2016) 029.
18. The ATLAS TDAQ Collaboration, *The ATLAS Data Acquisition and High Level Trigger system* 2016 JINST 11 P06008.
19. The CMS Collaboration, *The CMS Trigger*, 2017 JINST 12 P01020.
20. W. H. Smith, *Triggering at the LHC,* Ann. Rev. Nucl. Part. Sci. 66, 123 (2016).
21. CMS collaboration, Triggering on electrons, jets and tau leptons with the CMS upgraded calorimeter trigger for the LHC RUN II, 2016 JINST 11 C02008.
22. ATLAS Collaboration, *ATLAS Phase-II Upgrade Scoping Document*, CERN-LHCC-2015-020; CMS Collaboration, *Technical Proposal for the Phase-II Upgrade of the CMS Detector*, CERN-LHCC-2015-010.
23. M. zur Nedden (on behalf of the ATLAS Collaboration), *The LHC Run 2 ATLAS trigger system: design, performance and plans,* 2017 JINST 12 C03024.
24. M. Krivda et al. (The ALICE Collaboration), *The ALICE Central Trigger Processor (CTP) Upgrade* 2016 JINST 11 C03051.
25. Asbah, N. & ATLAS collaboration, *A hardware fast tracker for the ATLAS trigger,* Phys. Part. Nuclei Lett. (2016) 13: 527.
26. ALICE Collaboration, *Upgrade of the ALICE Readout & Trigger System,* CERN-LHCC-2013-019; M. Krivda and J. Pospisil for the ALICE Collaboration, *The ALICE Central Trigger Processor (CTP) Upgrade,* JINST 11, C03051 (2016); F Costa et al., *The detector read-out in ALICE during Run 3 and 4,* J. Phys.: Conf. Ser. 898 032011 (2017).

27. LHCb Collaboration, LHCb Trigger and Online Upgrade Technical Design Report, CERN-LHCC-2014-016; T. Szumlak, Real time analysis with the upgraded LHCb trigger in Run III, J. Phys.: Conf. Ser. 898 032051 (2017).
28. Apollinari, G. (ed.) et al., HL-LHC Preliminary Design Report, CERN-2015-005; P. Campana, M. Klute and P.S. Wells, *Physics Goals and Experimental Challenges of the Proton–Proton High-Luminosity Operation of the LHC,* Ann. Rev. Nucl. Part. Sci. 66, 273 (2016).
29. ATLAS Collaboration, *ATLAS Phase-II Upgrade Scoping Document*, CERN-LHCC-2015-20 (2015).
30. CMS Collaboration, The Phase-2 Upgrade of the CMS Level-2 Trigger, CERN-LHCC-2017-013 (2017); CMS Phase-II Upgrade Scope Document, CERN-LHCC-2015-19 (2015).
31. N.P. Ghanathe, et al., *Software and firmware co-development using high-level synthesis,* JINST 12 C01083 (2017).
32. S. Gorbunov et al., *ALICE HLT High Speed Tracking on GPU*, in IEEE Transactions on Nuclear Science, 58, no. 4, 1845 (2011).
33. C. Färber, R. Schwemmer, J. Machen and N. Neufeld, *Particle identification on an FPGA accelerated compute platform for the LHCb upgrade,* 2016 IEEE-NPSS Real Time Conference (RT), Padua, 2016, pp. 1–2.
34. S. Farrell et al., *The HEP.TrkX Project: deep neural networks for HL-LHC online and offline tracking,* EPJ Web of Conferences 150, 00003 (2017).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 13
Pattern Recognition and Reconstruction

R. Frühwirth, E. Brondolin, and A. Strandlie

13.1 Track Reconstruction

13.1.1 Introduction

Track reconstruction is the task of finding and estimating the trajectory of a charged particle, usually embedded in a static magnetic field to determine its momentum and charge. It involves pattern recognition algorithms and statistical estimation methods. Depending on the physics goals, not all charged tracks have to be reconstructed. For instance, in many cases there is a physically motivated lower limit on the momentum or transverse momentum of the particles to be found. Other examples are short-range secondary particles, such as δ-electrons, that normally need not be reconstructed. It may also be useful to reconstruct electron-positron pairs from photon conversions in order to check the distribution of material in the detector. Track reconstruction frequently proceeds in several steps:

1. Pattern recognition or Track finding: Finds the detector signals (hits) that are generated by the same charged particle.
2. Track fitting: Estimates for each track candidate the track parameters and the associated covariance matrix.
3. Test of track hypothesis: Tests for each track candidate whether all hits do indeed belong to the track and identifies outliers.

R. Frühwirth (✉) · E. Brondolin
Institute of High Energy Physics of the Austrian Academy of Sciences, Vienna, Austria
e-mail: Rudolf.Fruehwirth@oeaw.ac.at; erica.brondolin@cern.ch

A. Strandlie
NTNU—Norwegian University of Science and Technology, Gjøvik, Norway
e-mail: are.strandlie@ntnu.no

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_13

There are many different algorithms for track finding. A selection of them is described in Sects. 13.1.2.1 and 13.1.2.2. For an extended treatment of the subject, containing many examples, see the excellent exposition in [1]. The track fit takes a track candidate and estimates the track parameters (location, direction, momentum or curvature, see Sect. 13.1.3.2) from the detector hits (Sect. 13.1.3.4), taking into account the equation of motion (Sect. 13.1.3.2) in the magnetic field (Sect. 13.1.3.1) and the effects of the detector material on the trajectory (Sect. 13.1.3.3). In the test stage (Sect. 13.1.3.5) outliers are identified, i.e., hits which apparently do not belong to the track. If outliers are expected, the estimation procedure should be robust so that the estimated track is not significantly biased by the outliers. Some robust methods are discussed in Sect. 13.1.3.5. Section 13.1.4 treats track-based alignment, and Sect. 13.1.5 contains many useful formulas for determining the approximate momentum resolution of a tracking detector without extensive simulations.

13.1.2 Pattern Recognition

Pattern recognition or track finding methods can be divided into global and local methods. In a global method, all detector hits are treated on an equal footing, and all track candidates are found in parallel; in a local method, there is a privileged subset of hits which is used to find initial track candidates, which are then completed to full track candidates.

13.1.2.1 Global Methods

Typical global methods of track finding find the tracks in parallel, for instance by identifying peaks in a one- or two-dimensional histogram, or by observing the final state of a recurrent neural network.

Conformal Mapping

A popular method for finding circular particle tracks is the conformal mapping method [2]. It uses the fact that the mapping

$$u = \frac{x}{x^2 + y^2}, \quad v = \frac{y}{x^2 + y^2},$$

transforms circles going through the origin of an x–y coordinate system into straight lines of the form

$$v = \frac{1}{2b} - u\frac{a}{b},$$

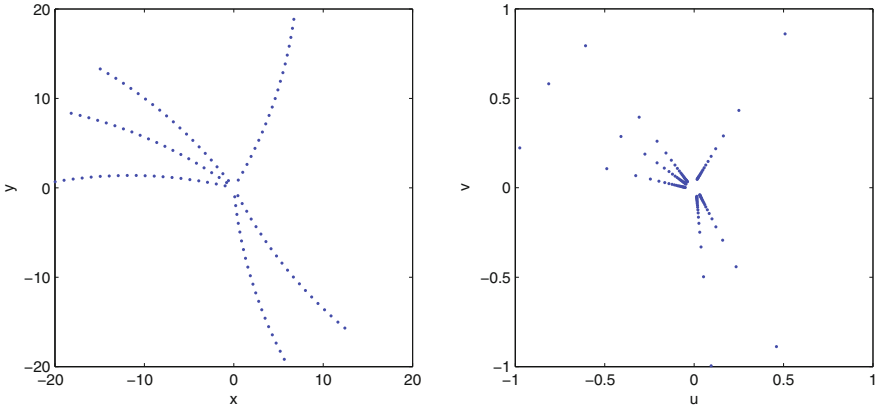


Fig. 13.1 The original measurements (left) and the transformed measurements (right) of six circular tracks

where the parameters a and b are defined by the circle equation

$$(x-a)^2 + (y-b)^2 = R^2 = a^2 + b^2.$$

The distance of the line to the origin is equal to $1/(2R)$, so that for large radius R it passes very close to the origin. The lines can be found by a histogramming method. After transforming the measurements in the u–v plane to polar coordinates and collecting the polar angle θ in a histogram, measurements belonging to the same particle will tend to create peaks in the histogram.

As an example, the measured points of six circular tracks are shown in the left hand panel of Fig. 13.1, while the transformed measurements are shown in the right hand panel of Fig. 13.1. The resulting histogram of the polar angle θ is shown in Fig. 13.2.

Hough Transform

In the general case of lines not passing close to the origin, a more general approach is needed in order to find the lines. A very popular method for this purpose is the Hough transform [3]. The principle of the Hough transform can be explained by noting that a straight line in an x–y coordinate system, $y = cx + d$, can also be regarded as a straight line in a c–d coordinate system by the transformation $d = -xc + y$. For a fixed point (x, y), the line in c–d space (also denoted parameter space) corresponds to all possible lines going through this point in x–y space (also denoted image space). Measurements lying along a straight line in image space therefore transform into lines in parameter space which cross at the specific value of the parameters of the line under consideration in image space. In practice, parameter space is discretized, and each measurement (x, y) leads to

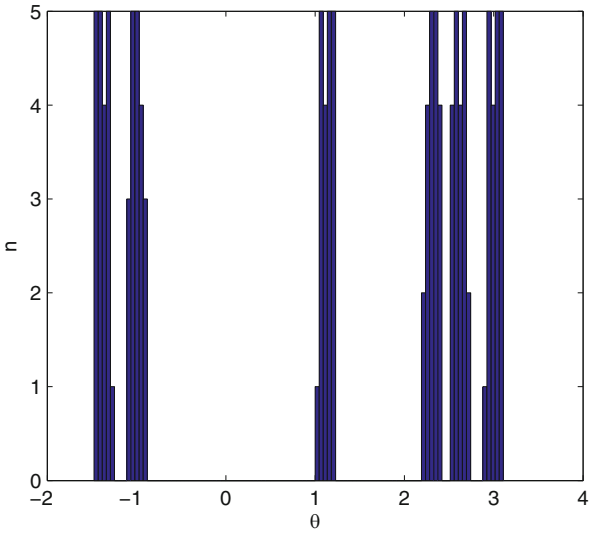


Fig. 13.2 Histogram of $\theta = \arctan(v/u)$

an increment of a set of histogram bins. Measurements lying along straight lines tend to create peaks in the histogram, and the lines can be found by searching for peaks in this histogram. The granularity of the discretization has to be optimized for each specific application, as it depends on the amount of noise present and the actual values of measurement uncertainties. A too fine-grained histogram can split or destroy peaks if the measurement uncertainties are non-negligible. On the other hand, a too course-grained histogram increases the sensitivity to noise, and nearby tracks may merge into a single peak.

The basic formulation of the Hough transform is an example of a divergent transform, i.e., one measurement in image space corresponds to a set of increments of histogram entries in parameter space. The Hough transform can also be made convergent by considering instead a pair of measurements in image space. A unique line passes through any such pair, and only one entry in the parameter space histogram needs to be incremented. A possible disadvantage of such an approach is that the number of pairs grows quadratically with the number of measurements in image space. In order to reduce computational complexity, one may consider only a randomly selected subset of all the pairs. This is the basic feature of probabilistic Hough transforms [4].

The Hough transform has turned out to be successful also for finding circles passing through the origin. With this constraint, two parameters are enough to uniquely describe the circle, and the task again amounts to finding peaks in a two-dimensional histogram. With three or more parameters, one has to search for clusters in multi-dimensional spaces, and in this case the Hough transform is in general less powerful than in the two-dimensional case.

For track finding in drift tubes, with their inherent left-right ambiguity, the drift circles can be transformed to sine curves in the (r, θ) space by applying a Legendre transform [5]. The peaks at the intersections of several sine curves represent the common tangents to a set of several circles.

Neural Networks

Recurrent neural networks of the Hopfield type [6] are used in finding solutions to certain kinds of combinatorial optimization problems, i.e., problems that can be formulated as finding the minimum of an energy function

$$E = -\frac{1}{2} \sum_i \sum_j T_{ij} S_i S_j$$

with respect to the configuration of n binary-valued neurons $S_i, i = 1, \ldots, n$ and fixed connection weights $T_{ij}, i, j = 1, \ldots, n$. It was realized independently in [7] and [8] that the track finding problem can be formulated as a minimization problem of this kind. The neurons are links between measurements which potentially belong to the same track. The connection weights T_{ij} have a structure which favors links sharing a measurement and pointing in a similar direction. The standard network dynamics leads to a solution corresponding to a local minimum of the energy function. A better solution is to apply a mean-field annealing technique [9], which introduces a temperature parameter and thereby allows the neurons to take all values in the interval between the two original binary values. The network is initialized at a high temperature, the mean-field equations are iteratively solved as the network is cooled down, and the low-temperature limit is taken in the end. At a significantly lower computational effort, the approximate solutions found by the mean-field technique have been shown to be very close to the exact solutions [10]. For applications of the Hopfield network in experiments see e.g. [11–15].

The energy function of the Hopfield network can be generalized in order to take into account the track model (see Sect. 13.1.3.2), i.e., the known parametric form of the tracks. The resulting algorithm is called elastic tracking or elastic arms [16–19]. A related generalization is the elastic net, originally used to tackle the traveling salesman problem [20]. Applications to track finding are described in [21] and [22, 23].

13.1.2.2 Local Methods

A local track finding method finds the tracks sequentially, starting from an initial track segment or an initial collection of measured points.

Track Road

The track road method starts out with a set of measurements that potentially belong to the same track, typically one close to the vertex area, one far out in the tracking detector, and one in the middle. The track model can then be used, either exactly or approximately, if speed is an important issue, to interpolate between the measurements and create a road around the hypothesized track. Measurements inside the road are then collected. The number of measurements and the quality of the subsequent track fit are used to determine whether the track candidate should be kept or discarded.

Track Following

A track following procedure takes a track seed as a starting point. A seed is often a short track segment, potentially including a constraint of the position of the vertex region. Seeds can be generated at the inner part of the tracking detector, where the measurements frequently are of very high precision, or at the outer part, where the track density is lower. From the seed, the track is extrapolated to the next detector unit. As for the track roads method, this can be done either with the full track model or with an approximate, simplified model. The measurement closest to the predicted track is included in the track candidate, and the track is extrapolated again.

Kalman Filter

The Kalman filter [24–26] can be regarded as a statistically optimal track following procedure. It works by alternating prediction and update steps. Starting from the seed, the track parameters and their covariance matrix are extrapolated to the next detector unit containing a measurement, using the full track model. If the measurement is compatible with the prediction, it is included in the track candidate, and the track parameters and their covariance matrix are updated with the information from the measurement. The procedure is repeated until too many detector units without compatible measurements are traversed or the end of the tracking detector is reached.

In the original formulation of the method, the measurement closest to the predicted track is included in the track candidate [27]. However, if the density of measurements is high, the closest measurement might originate from another particle or from noise in the detector electronics. Including the wrong measurement could therefore lead to a wrong subsequent prediction and ultimately to the loss of the track. The currently most popular approach, the combinatorial Kalman filter, avoids such losses by splitting the track candidate into several branches when several compatible measurements are found after the prediction [28]. In order to take into account detector inefficiencies an additional branch with a missing hit can be generated.

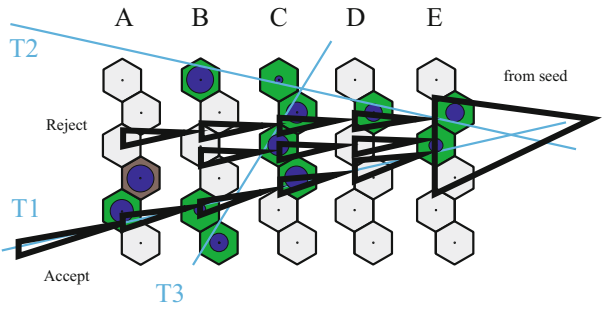


Fig. 13.3 An example of the combinatorial Kalman filter (reprinted from R. Mankel [28], with permission from Elsevier)

All branches are extrapolated to the next detector layer containing compatible measurements. A branch is split again if several measurements are compatible with the branch prediction. Branches are removed if too many detector units without compatible measurements are traversed or if the quality of the track candidate, in terms of the value of a χ^2 statistic, is too low. If there are several surviving candidates after the end of the detector has been reached, the candidate with most measurements and the lowest value of the χ^2 statistic is kept and regarded as the final track candidate. An example is shown in Fig. 13.3.

A similar track finding method has been formulated in the language of cellular automata [23, 29]. The combinatorial problem can also be solved by using generalized, adaptive versions of the Kalman filter [30–32].

13.1.3 Estimation of Track Parameters

13.1.3.1 Magnetic Field Representation

The presence of a magnetic field in a tracking detector causes a bending of the trajectory of a charged particle, and, hence, allows a measurement of the particle momentum. A precise knowledge of the magnetic field is therefore crucial for accurate estimates of the particle momenta.

The magnetic field can be calculated by solving Maxwell's equations, knowing the detailed configuration of the current sources and the magnetic materials in the detector volume. In the general case, a numerical solution of these equations in terms of a finite-element analysis is needed. In special cases, the field can be found by less general approaches. The simplest situation is a solenoidal magnet, providing a homogeneous field in a large volume. Also, it is known that the field inside a volume with no magnetic material can be determined by knowledge of the field on the volume boundary only [33]. Measurements of the field on the volume boundary allows an estimation of coefficients of polynomials obeying Maxwell's

equations. Field measurements inside the volume are used to evaluate the quality of the calculated field. If the measurements inside the volume are precise enough, they can be used to further refine the knowledge of the field by being included in the estimation procedure of the abovementioned coefficients [34].

In a track reconstruction application, fast access to the value of the magnetic field at any point inside the detector volume is crucial. For this purpose, a numerical representation of the field is needed. A frequently used approach is to create a table of the magnetic field values at a grid of points and to determine the field at points between the grid nodes by linear or quadratic interpolation. An alternative approach is to divide the detector volume into several sub-volumes and to fit the coefficients of low-order polynomials to the known field values inside each sub-volume [35, 36]. If the number of sub-volumes is large, potentially many coefficients have to be determined. On the other hand, once the coefficients are determined, the field access is very fast. Also, the derivatives of the field, which are needed by some track reconstruction algorithms, can be computed as fast as the field itself.

13.1.3.2 Track Models

Consider a charged particle with mass m and charge $Q = qe$, e being the elementary charge. Its trajectory $\boldsymbol{x}(t)$ in a magnetic field $\boldsymbol{B}(\boldsymbol{x})$ is determined by the equations of motion given by the Lorentz force $\boldsymbol{F} \propto q\boldsymbol{v} \times \boldsymbol{B}$, where $\boldsymbol{v} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$ is the velocity of the particle. In vacuum, Newton's second law reads [37]

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = kq\boldsymbol{v}(t) \times \boldsymbol{B}(\boldsymbol{x}(t)), \tag{13.1}$$

where $\boldsymbol{p} = \gamma m\boldsymbol{v}$ is the momentum of the particle, $\gamma = (1 - \boldsymbol{v}^2/c^2)^{-1/2}$ is the Lorentz factor, and k is a unit-dependent proportionality factor. If $\boldsymbol{p}$ is in GeV/c, $\boldsymbol{x}$ is in meters, and $\boldsymbol{B}$ is in Tesla, $k = 0.29979\ \mathrm{GeV}/c\ \mathrm{T}^{-1}\,\mathrm{m}^{-1}$. The trajectory is uniquely defined by the initial conditions, the six degrees of freedom specified for instance by the initial position and the initial velocity. If these are tied to a surface, five degrees of freedom are necessary and sufficient. Geometrical quantities other than position and velocity can also be used to specify the initial conditions. The collection $\boldsymbol{q}$ of these quantities is called the initial track parameters or the initial state vector.

Equation (13.1) can be written in terms of the path length $s(t)$ along the trajectory instead of t, giving [37]

$$\frac{\mathrm{d}^2\boldsymbol{x}}{\mathrm{d}s^2} = \frac{kq}{|\boldsymbol{p}|} \cdot \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s} \times \boldsymbol{B}(\boldsymbol{x}(s)) = F(s, \boldsymbol{x}(s), \dot{\boldsymbol{x}}(s)). \tag{13.2}$$

In simple situations this equation has analytical solutions. In a homogeneous magnetic field the trajectory is a helix; it reduces to a straight line in the limit of a vanishing field. In the general case of an inhomogeneous field, numerical

methods can be used, such as Runge–Kutta integration of the equations of motion or parametrization by polynomials or splines [37]. Among Runge–Kutta methods, the Runge-Kutta-Nyström algorithm is specially designed for second-order equations such as Eq. (13.2). In the fourth-order version a step of length h, starting at $s = s_n$, is computed by [37]

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + h\dot{\boldsymbol{x}}_n + h^2(k_1 + k_2 + k_3)/6, \quad \dot{\boldsymbol{x}}_{n+1} = \dot{\boldsymbol{x}}_n + h(k_1 + 2k_2 + 2k_3 + k_4)/6,$$

with

$$\begin{aligned} k_1 &= F(s_n, \boldsymbol{x}_n, \dot{\boldsymbol{x}}_n), \\ k_2 &= F(s_n + h/2, \boldsymbol{x}_n + h\dot{\boldsymbol{x}}_n/2 + h^2 k_1/8, \dot{\boldsymbol{x}}_n + hk_1/2), \\ k_3 &= F(s_n + h/2, \boldsymbol{x}_n + h\dot{\boldsymbol{x}}_n/2 + h^2 k_1/8, \dot{\boldsymbol{x}}_n + hk_2/2), \\ k_4 &= F(s_n + h, \boldsymbol{x}_n + h\dot{\boldsymbol{x}}_n + h^2 k_3/2, \dot{\boldsymbol{x}}_n + hk_3), \end{aligned}$$

where $\boldsymbol{x}_n$ is the position of the particle at $s = s_n$ and $\dot{\boldsymbol{x}}_n$ is the unit tangent vector. The magnetic field needs to be looked up for the calculation of k_2, k_3 and k_4, i.e., three times per step. If the field at the final position $\boldsymbol{x}_{n+1}$, which is the starting position of the next step, is approximated by the field used for k_4, only two lookups are required per step. If the field is (almost) homogeneous, as for example in a solenoid, the step size h can be chosen to be constant; otherwise a variable step size is more efficient. The step size can be optimized using an adaptive version of the Runge-Kutta-Nyström algorithm [38]. Note that the error of a step of length h may be larger than $O(h^5)$ if the magnetic field does not have smooth derivatives, as is the case if it is computed by linear interpolation. If the field is represented by low-order polynomials in sub-volumes, Runge–Kutta steps should terminate at volume boundaries.

Different detector geometries often lead to different choices of the parametrization. However, the parametrization of the trajectory should comply to some basic requirements: the parameters should be continuous with respect to small changes of the trajectory; the choice of track parameters should facilitate the local expansion of the track model into a linear function; and the uncertainties of the estimated values of the parameters should follow a Gaussian distribution as closely as possible. For example, curvature should be used rather than radius of curvature, and inverse (transverse) momentum rather than (transverse) momentum.

The track model, given by the solution of the equations of motion, describes how the state vector $\boldsymbol{q}_k$ at a given surface k depends on the state vector at a different surface i:

$$\boldsymbol{q}_k = \boldsymbol{f}_{k|i}(\boldsymbol{q}_i),$$

where $\boldsymbol{f}_{k|i}$ is the track propagator from surface i to surface k. When analytical solutions of the equations of motion exist, the track propagator is also analytical. Even in

a homogeneous magnetic field, the path length can be determined analytically only for propagation to cylinders with symmetry axis parallel to the field direction or to planes orthogonal to the field direction. Otherwise, a Newton iteration or a parabolic approximation has to be used to find the path length.

For track reconstruction purposes, the covariance matrix of the estimated track parameters needs to be propagated along with the track parameters themselves. The track propagator is often a non-linear function of the track parameters at the initial surface, but the covariance matrix has to be transported under the assumption of a linear track model. This procedure, called linear error propagation, is based on a Taylor expansion of the track propagator, keeping only first-order terms. These first-order terms, defining the Jacobians of the track model, are given by

$$\boldsymbol{F}_{k|i} = \left.\frac{\partial \boldsymbol{q}_k}{\partial \boldsymbol{q}_i}\right|_{\check{\boldsymbol{q}}_i},$$

where $\check{\boldsymbol{q}}_i$ is the expansion point in surface i. For analytical track models, the Jacobian is also analytical. If, for example, the magnetic field is homogeneous, the general case of propagation to a plane of arbitrary spatial orientation uses a curvilinear coordinate frame moving along with the trajectory as a means of deriving the required Jacobians [39].

In the general case of a non-analytical track model, the Jacobians cannot be computed analytically either. The most straightforward approach is to calculate the relevant derivatives in a purely numerical way. The basis for these calculations is a reference trajectory corresponding to the expansion point. In addition, five other trajectories are created, corresponding to small variations in each of the track parameters. By propagating these five trajectories to the destination surface, numerical derivatives can be obtained. A potential disadvantage of such an approach is its computational complexity, as six trajectories have to be propagated instead of a single one. Much less computational load is introduced by transporting the Jacobian terms in parallel to the track parameters during the Runge–Kutta integration [40, 41], avoiding the need for propagating auxiliary trajectories.

The measurement model describes the functional dependence of the measured quantities on the state vector at a detector surface k:

$$\boldsymbol{m}_k = \boldsymbol{h}_k(\boldsymbol{q}_k).$$

The vector of measurements $\boldsymbol{m}_k$ usually contains the measured coordinates, but may contain also other quantities, e.g. measurements of direction or even momentum. In a pixel detector or in a double-sided silicon strip detector, $\boldsymbol{m}_k$ is two-dimensional; in a one-sided strip detector, it is one-dimensional. In a drift chamber or a multi-wire proportional chamber with several layers, the measurement may be a track segment resulting from an internal track reconstruction. In this case the vector $\boldsymbol{m}_k$ may be four- or five-dimensional, depending on whether the curvature can be estimated or not.

In most cases the function $\boldsymbol{h}_k(\boldsymbol{q}_k)$ includes a transformation of the state vector $\boldsymbol{q}_k$ into the local coordinate system of the detector. For use in track reconstruction, the Jacobian of this transformation is needed:

$$\boldsymbol{H}_k = \left.\frac{\partial \boldsymbol{m}_k}{\partial \boldsymbol{q}_k}\right|_{\check{\boldsymbol{q}}_k}, \tag{13.3}$$

where $\check{\boldsymbol{q}}_k$ is the expansion point in surface k. In many cases the Jacobian contains only rotations and projections, and thus can be computed analytically.

The measurement is always smeared by a measurement error:

$$\boldsymbol{m}_k = \boldsymbol{h}_k(\boldsymbol{q}_k) + \boldsymbol{\varepsilon}_k.$$

The mean value and the covariance matrix of $\boldsymbol{\varepsilon}_k$ depend on the detector type and the detector geometry and have therefore in general to be calibrated for each detector unit independently. The measurement error is often assumed to follow a Gaussian distribution, but frequently exhibits tails which are incompatible with this assumption. In this case a Gaussian mixture is a more appropriate model.

13.1.3.3 Material Effects

A charged particle crossing a tracking detector interacts with the material of the detector. The most important types of interactions in track reconstruction are multiple Coulomb scattering, energy loss by ionization, and energy loss by bremsstrahlung. For an in-depth treatment of material effects see Chapter 2.

Multiple Coulomb Scattering

Elastic Coulomb scattering of particles heavier than the electron is dominated by the atomic nucleus. For small angles the differential cross-section is approximately equal to

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\theta} = 2\pi \left(\frac{2Ze^2}{pv}\right)^2 \frac{1}{\theta^3},$$

where θ is the polar angle of the scattering, Z is the charge of the nucleus in units of the elementary charge e, v is the velocity of the scattered particle, and p is its momentum [42]. Because of screening effects and the finite size of the nucleus the differential cross-section is modified to [43]

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\theta} = 2\pi \left(\frac{2Ze^2}{pv}\right)^2 \frac{\theta}{(\theta^2 + \theta_{\min}^2)^2}, \quad 0 \le \theta \le \theta_{\max}.$$

If the momentum p is given in GeV/c, the lower and upper limits are approximately equal to

$$\theta_{\min} \approx \frac{2.66 \cdot 10^{-6} Z^{1/3}}{p}, \quad \theta_{\max} \approx \frac{0.14}{A^{1/3} p}.$$

The average number of scattering processes in a layer of thickness d (in cm) is given by

$$N(d) = d\sigma \frac{N_{\mathrm{A}} \rho}{A},$$

where σ is the integrated elastic cross section, N_{A} is the Avogadro constant, ρ is the density of the material (in g/cm^3), and A is the atomic mass of the nucleus. In track reconstruction it is convenient to work with the projected scattering angles in two perpendicular planes. The projected multiple scattering angle θ_{P} is equal to the sum of the projected single scattering angles, and its variance can be obtained by multiplying the variance of the projected single scattering angle by the average number of scatters, the projected single scattering angles being uncorrelated. With increasing thickness d the distribution of the projected scattering angle approaches a normal distribution, and the two projected angles become independent. For thin scatterers, however, the width of the Gaussian core is notably narrower than is indicated by the variance [42]. This is taken into account by Highland's formula for the standard deviation of the projected scattering angle [44]:

$$\sigma_{\mathrm{P}} = E(\theta_{\mathrm{P}}^2)^{1/2} = \frac{0.0136}{\beta p} \sqrt{d/X_0} \left[1 + 0.038 \ln(d/X_0)\right],$$

where X_0 is the radiation length of the material in cm, $\beta = v/c$ is the particle velocity in units of c, and p is the particle momentum in GeV/c. The logarithmic correction ceases to be applicable above $d \approx X_0$.

If a scatterer is sufficiently thin, the transverse offset of the track due to multiple scattering can be neglected. Only the track direction is affected in this case. If the direction is represented by the polar angle θ and the azimuthal angle φ, their joint covariance matrix is given by

$$\mathrm{var}(\Delta\theta) = \sigma_{\mathrm{P}}^2, \quad \mathrm{var}(\Delta\varphi) = \sigma_{\mathrm{P}}^2 / \sin^2\theta, \quad \mathrm{cov}(\Delta\theta, \Delta\varphi) = 0.$$

If the direction is represented by the direction tangents $t_x = \mathrm{d}x/\mathrm{d}z$ and $t_y = \mathrm{d}y/\mathrm{d}z$, the covariance matrix is [45]

$$\mathrm{Var}[(\Delta t_x, \Delta t_y)^{\mathrm{T}}] = \sigma_P^2 (1 + t_x^2 + t_y^2) \begin{pmatrix} 1 + t_x^2 & t_x t_y \\ t_x t_y & 1 + t_y^2 \end{pmatrix}.$$

If the direction is represented by the direction cosines $c_x = \mathrm{d}x/\mathrm{d}s$ and $c_y = \mathrm{d}y/\mathrm{d}s$, the covariance matrix is [45]

$$\mathrm{Var}[(\Delta c_x, \Delta c_y)^{\mathrm{T}}] = \sigma_P^2 \begin{pmatrix} (1-c_x)^2 & -c_x c_y \\ -c_x c_y & (1-c_y)^2 \end{pmatrix}.$$

In all cases the projected variance σ_{P}^2 takes into account the effective amount of material crossed by the track.

If the transverse offset cannot be neglected, its variance and its correlation with the angle have to be taken into account. Assume that the particle passes a scatterer of length d, traveling along the z-axis. Neglecting the curvature of the track, the joint covariance matrix of the offset Δx and the scattering angle θ_x in the x–z projection is

$$\mathrm{var}(\Delta x) = \sigma_0^2 d^3/3, \quad \mathrm{var}(\theta_x) = \sigma_0^2 d, \quad \mathrm{cov}(\Delta x, \theta_x) = \sigma_0^2 d^2/2,$$

where σ_0^2 is the variance of the projected scattering angle per unit length. If the particle enters the scatterer at $z = 0$ with direction (t_x, t_y), the joint covariance matrix of the offsets Δx, Δy and the angles Δt_x, Δt_y at $z = d$ is

$$\mathrm{Var}\left[\begin{pmatrix} \Delta x \\ \Delta y \\ \Delta t_x \\ \Delta t_y \end{pmatrix}\right] = \sigma_0^2 (1+t_x^2+t_y^2) \begin{pmatrix} (1+t_x^2)D^3/3 & t_x t_y D^3/3 & (1+t_x^2)D^2/2 & t_x t_y D^2/2 \\ t_x t_y D^3/3 & (1+t_y^2)D^3/3 & t_x t_y D^2/2 & (1+t_y^2)D^2/2 \\ (1+t_x^2)D^2/2 & t_x t_y D^2/2 & (1+t_x^2)D & t_x t_y D \\ t_x t_y D^2/2 & (1+t_y^2)D^2/2 & t_x t_y D & (1+t_y^2)D \end{pmatrix},$$

where $D = (1 + t_x^2 + t_y^2)^{1/2} d$ is the effective thickness crossed. If the direction is represented by θ and φ, the covariance matrix can be computed via the transformation

$$t_x = \tan(\theta)\cos(\varphi), \quad t_y = \tan(\theta)\sin(\varphi),$$

and linear error propagation with the Jacobian

$$\boldsymbol{T} = \frac{\partial(\theta, \varphi)}{\partial(t_x, t_y)} = \begin{pmatrix} \frac{t_x}{\sqrt{t_x^2+t_y^2}(1+t_x^2+t_y^2)} & \frac{t_y}{\sqrt{t_x^2+t_y^2}(1+t_x^2+t_y^2)} \\ -\frac{t_y}{t_x^2+t_y^2} & \frac{t_x}{t_x^2+t_y^2} \end{pmatrix} = \begin{pmatrix} \frac{\cos(\varphi)}{1+\tan^2(\theta)} & \frac{\sin(\varphi)}{1+\tan^2(\theta)} \\ -\frac{\sin(\varphi)}{\tan(\theta)} & \frac{\cos(\varphi)}{\tan(\theta)} \end{pmatrix}.$$

For analogous formulas in cylindrical coordinates, see [46].

If the curvature of the track cannot be neglected, the simplest approach is a stepwise integration of the equation of motion, assuming the validity of a helical track model within each step and considering each such step as a thin scatterer [39, 47].

Energy Loss

For particles other than electrons the energy loss in material is almost exclusively due to scattering on electrons. The momentum correction Δp in a material layer of thickness d is calculated by integrating the Bethe-Bloch formula [37]:

$$\Delta p = \int_0^d \frac{\mathrm{d}p}{\mathrm{d}x}\mathrm{d}x = \int_0^d \frac{1}{\beta}\frac{\mathrm{d}E}{\mathrm{d}x}\mathrm{d}x = \int_0^d \frac{K}{\beta^3}\left[\ln\frac{2m_\mathrm{e}c^2\beta^2\gamma^2}{\langle I\rangle} - \beta^2\right]\mathrm{d}x, \qquad (13.4)$$

where K is a constant depending on the material, m_e is the electron mass, $\langle I\rangle$ is the average ionization potential of the material, and $\beta = v/c$ and $\gamma = E/mc^2$ are the usual kinematic parameters. The ratio $\langle I\rangle/Z$ is about 20 eV for hydrogen and helium, between 12 and 16 eV for light nuclei, and around 10 eV for heavy nuclei [44]. For practical purposes, the differential energy loss $\mathrm{d}E/\mathrm{d}x$ is a function only of β. For small β, it decreases like $1/\beta^2$. It has a minimum, the position of which drops with increasing Z from $\beta\gamma \approx 3.5$ (carbon) to $\beta\gamma \approx 3$ (lead). In terms of momentum, the minimum is at $p = \beta\gamma mc$ and thus depends on the mass of the particle. This dependency is used for particle identification. The energy loss at the minimum can be parameterized for $Z \geq 6$ by [44]:

$$(\mathrm{d}E/\mathrm{d}x)_\mathrm{min} = (2.35 - 0.64\ln_{10} Z)\ \mathrm{MeV\,g^{-1}cm^2}.$$

From this the constant K in Eq. (13.4) can be calculated. For large $\beta\gamma$ the energy loss increases like $\ln(\beta\gamma)$; this is called the relativistic rise. For momenta in the vicinity of the minimum $\mathrm{d}E/\mathrm{d}x$ can be considered as constant, giving $\Delta p \approx (\mathrm{d}E/\mathrm{d}x)_\mathrm{min} \cdot d\rho/\beta$, ρ being the density of the material.

Bremsstrahlung

For an electron (or positron) passing through matter the most significant contribution to energy loss is bremsstrahlung, the emission of photons in the electric field of an atomic nucleus. In the Bethe–Heitler model [48] the relative energy loss is distributed independently of the energy. Let d be the path length in the material in units of radiation length, and z the fraction of energy remaining after the material is traversed. Then the distribution of z is given by the following probability density function:

$$f(z) = \frac{(-\ln z)^{c-1}}{\Gamma(c)}, \quad 0 \leq z \leq 1,$$

where $\Gamma(x)$ is Euler's gamma function and $c = d/\ln 2$. For high energy electrons $p \approx E$, so the momentum correction is $\Delta p \approx p(z-1)$. The first two moments of Δp are

$$E(\Delta p) = p(2^{-c} - 1), \quad \mathrm{var}(\Delta p) = p^2(3^{-c} - 4^{-c}).$$

The moments can be used for a Gaussian representation of bremsstrahlung as an additional process noise in the Kalman filter (see Sect. 13.1.3.4). As this is a very crude approximation, more sophisticated methods have been developed that take into account the actual shape of the distribution. One of them is the Gaussian-sum filter [49, 50], see Sect. 13.1.3.5. A computationally less intensive approach is described in [51].

13.1.3.4 Estimation Methods

The main task of the track fit is to estimate the values of a set of parameters describing the state of a particle somewhere in the detector, often at a reference surface close to the interaction vertex. The information from the measurements created by the particle while traversing the tracking detector should be processed in an optimal manner. If the track model is truly linear, i.e., if the measurements are strictly linear functions of the track parameters, and all stochastic disturbances entering the estimation procedure are Gaussian, the linear least-squares method is the optimal one [37]. Since track parameter propagation in general is a nonlinear procedure, strict linearity holds very rarely in practice. The relation between the track parameter vector $\boldsymbol{q}_0$ at a reference surface and the measurement vector $\boldsymbol{m}_k$ at a detector layer k is a function $\boldsymbol{d}_k$ given by

$$\boldsymbol{m}_k = \boldsymbol{d}_k(\boldsymbol{q}_0) + \boldsymbol{\gamma}_k,$$

where $\boldsymbol{\gamma}_k$ is a noise term containing the measurement error of $\boldsymbol{m}_k$ and all multiple scattering in front of $\boldsymbol{m}_k$. The function $\boldsymbol{d}_k$ is a composition of the measurement model function $\boldsymbol{h}_k$ and the track propagator functions $\boldsymbol{f}_{i|i-1}$ (see Sect. 13.1.3.2):

$$\boldsymbol{d}_k = \boldsymbol{h}_k \circ \boldsymbol{f}_{k|k-1} \circ \cdots \circ \boldsymbol{f}_{2|1} \circ \boldsymbol{f}_{1|0}.$$

For the linear least-squares method $\boldsymbol{d}_k$ has to be linearized around some expansion point, providing the Jacobian $\boldsymbol{D}_k$ of each $\boldsymbol{d}_k$:

$$\boldsymbol{D}_k = \boldsymbol{H}_k \boldsymbol{F}_{k|k-1} \cdots \boldsymbol{F}_{2|1} \boldsymbol{F}_{1|0},$$

with $\boldsymbol{H}_k$ from Eq. (13.3). The covariance matrix of $\boldsymbol{\gamma}_k$ is obtained by linear error propagation:

$$\mathrm{var}(\boldsymbol{\gamma}_k) = \boldsymbol{V}_k + \boldsymbol{H}_k(\boldsymbol{F}_{k|1}\boldsymbol{Q}_1\boldsymbol{F}_{k|1}{}^{\mathrm{T}} + \cdots + \boldsymbol{F}_{k|k-1}\boldsymbol{Q}_{k-1}\boldsymbol{F}_{k|k-1}{}^{\mathrm{T}} + \boldsymbol{Q}_k)\boldsymbol{H}_k{}^{\mathrm{T}},$$

where $\boldsymbol{V}_k$ is the covariance matrix of the measurement error $\boldsymbol{\varepsilon}_k$ of $\boldsymbol{m}_k$, and $\boldsymbol{Q}_j$ is the covariance matrix of multiple scattering after layer $j-1$ up to and including layer j. The part of $\boldsymbol{Q}_j$ originating from scattering between the layers has to be transported to layer j by the appropriate Jacobian. Because of the cumulative effect of multiple

scattering $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_k$ are correlated. If $i < k$, the covariance is given by

$$\mathrm{cov}(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_k) = \boldsymbol{H}_i(\boldsymbol{F}_{i|1}\,\boldsymbol{Q}_1\boldsymbol{F}_{k|1}{}^{\mathrm{T}} + \cdots + \boldsymbol{F}_{i|i-1}\,\boldsymbol{Q}_{i-1}\boldsymbol{F}_{k|i-1}{}^{\mathrm{T}} + \boldsymbol{Q}_i\boldsymbol{F}_{k|i}{}^{\mathrm{T}})\boldsymbol{H}_k{}^{\mathrm{T}}.$$

The observations $\boldsymbol{m}_k$, the functions $\boldsymbol{d}_k$, their Jacobians $\boldsymbol{D}_k$, and the noise $\boldsymbol{\gamma}_k$ are now collected in single vectors and a matrix:

$$\boldsymbol{m} = \begin{pmatrix} \boldsymbol{m}_1 \\ \cdot \\ \cdot \\ \boldsymbol{m}_n \end{pmatrix}, \quad \boldsymbol{d} = \begin{pmatrix} \boldsymbol{d}_1 \\ \cdot \\ \cdot \\ \boldsymbol{d}_n \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 \\ \cdot \\ \cdot \\ \boldsymbol{D}_n \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \cdot \\ \cdot \\ \boldsymbol{\gamma}_n \end{pmatrix},$$

where n is the total number of measurement layers. This gives the following model:

$$\boldsymbol{m} = \boldsymbol{d}(\boldsymbol{q}_0) + \boldsymbol{\gamma},$$

which now can be linearized into

$$\boldsymbol{m} = \boldsymbol{D}\boldsymbol{q}_0 + \boldsymbol{c} + \boldsymbol{\gamma},$$

where $\boldsymbol{c}$ is a constant vector. The global least-squares estimate of $\boldsymbol{q}_0$ is given by

$$\tilde{\boldsymbol{q}}_0 = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{G}\,(\boldsymbol{m} - \boldsymbol{c}),$$

where $\boldsymbol{V} = \boldsymbol{G}^{-1}$ is the non-diagonal covariance matrix of $\boldsymbol{\gamma}$. The quality of the initial expansion point can be monitored by using the obtained estimate as a new expansion point, and the state vector estimate can hence be re-calculated. Such a procedure is repeated until convergence, defined by a suitable stopping criterion.

If the track model is a circle and multiple scattering and energy loss can be neglected, the estimation can be simplified substantially. Explicit estimators are given in [52] for the center and radius of the circle, and in [53] for the curvature, the direction and the distance from a fixed point. Other algorithms are based on conformal mapping in the plane [2] or on a mapping to the Riemann sphere [54–56].

If there is strong multiple scattering, the estimated track can be quite far away from the real track. In order to follow the actual track more closely, two projected scattering angles can be explicitly estimated at each detector layer or at a set of virtual breakpoints inside a continuous scatterer [45, 57]. The breakpoint method, also known as General Broken Lines [58], and the global least-squares method are equivalent, as far as the estimate of the state vector $\boldsymbol{q}_0$ is concerned [59].

If the number of measurements or the number of breakpoints is substantial, the computational cost of these methods can be high due to the necessity of inverting large matrices during the estimation procedure. The Kalman filter, a recursive formulation of the least-squares method, requires the inversion of only

small matrices and exhibits the same attractive feature as the breakpoint method of following the actual track quite closely [26, 60].

As mentioned earlier, the Kalman filter proceeds by alternating prediction and update steps. The prediction step is the propagation of the track parameter vector from one detector layer containing a measurement to the next,

$$\boldsymbol{q}_{k|k-1} = \boldsymbol{f}_{k|k-1}(\boldsymbol{q}_{k-1|k-1}),$$

and the associated covariance matrix,

$$\boldsymbol{C}_{k|k-1} = \boldsymbol{F}_{k|k-1}\boldsymbol{C}_{k-1|k-1}{\boldsymbol{F}_{k|k-1}}^{\mathrm{T}} + \boldsymbol{Q}_k.$$

The update step is the correction of the predicted state vector due to the information from the measurement in layer k:

$$\boldsymbol{q}_{k|k} = \boldsymbol{q}_{k|k-1} + \boldsymbol{K}_k\left[\boldsymbol{m}_k - \boldsymbol{h}_k(\boldsymbol{q}_{k|k-1})\right],$$

where the gain matrix $\boldsymbol{K}_k$ is given by

$$\boldsymbol{K}_k = \boldsymbol{C}_{k|k-1}{\boldsymbol{H}_k}^{\mathrm{T}}\left(\boldsymbol{V}_k + \boldsymbol{H}_k\boldsymbol{C}_{k|k-1}{\boldsymbol{H}_k}^{\mathrm{T}}\right)^{-1}.$$

The update of the covariance matrix is given by

$$\boldsymbol{C}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\,\boldsymbol{C}_{k|k-1}.$$

The information filter is a mathematically equivalent, but numerically more stable formulation of the Kalman filter. In the information filter, the update of the state vector reads

$$\boldsymbol{q}_{k|k} = \boldsymbol{C}_{k|k}\left[\left(\boldsymbol{C}_{k|k-1}\right)^{-1}\boldsymbol{q}_{k|k-1} + {\boldsymbol{H}_k}^{\mathrm{T}}\boldsymbol{V}_k^{-1}\boldsymbol{m}_k\right],$$

whereas the update of the covariance matrix is given by

$$\boldsymbol{C}_{k|k} = \left[\left(\boldsymbol{C}_{k|k-1}\right)^{-1} + {\boldsymbol{H}_k}^{\mathrm{T}}\boldsymbol{V}_k^{-1}\boldsymbol{H}_k\right]^{-1}.$$

The implementation of the Kalman filter requires the computation of the Jacobians $\boldsymbol{F}_{k|k-1}$ and $\boldsymbol{H}_k$. A compilation of analytical formulas for two important cases (fixed-target configuration and solenoidal configuration) is given in [61].

Full information of the track parameters at the end of the track is obtained when all n measurements in the track candidate have been processed by the filter. The full information can be propagated back to all previous estimates by another iterative procedure, the Kalman smoother. A step of the smoother from layer $k+1$ to layer

k is for the state vector

$$\boldsymbol{q}_{k|n} = \boldsymbol{q}_{k|k} + \boldsymbol{A}_k(\boldsymbol{q}_{k+1|n} - \boldsymbol{q}_{k+1|k}),$$

where the smoother gain matrix is given by

$$\boldsymbol{A}_k = \boldsymbol{C}_{k|k}\boldsymbol{F}_{k+1|k}{}^{\mathrm{T}}(\boldsymbol{C}_{k+1|k})^{-1}.$$

The smoothed covariance matrix is

$$\boldsymbol{C}_{k|n} = \boldsymbol{C}_{k|k} - \boldsymbol{A}_k(\boldsymbol{C}_{k+1|k} - \boldsymbol{C}_{k+1|n})\boldsymbol{A}_k{}^{\mathrm{T}}.$$

The smoother can also be realized by combining two filters running in opposite directions: a forward filter from $\boldsymbol{m}_1$ to $\boldsymbol{m}_n$ and a backward filter from $\boldsymbol{m}_n$ to $\boldsymbol{m}_1$. The smoothed states are the weighted mean of the predicted states of one filter and the updated states of the other filter. This approach is numerically more stable than the gain matrix formulation of the smoother.

13.1.3.5 Track Quality and Robust Estimation

Robust estimators are insensitive to outliers, i.e., measurements that are biased or do not originate from the particle creating the majority of the hits in a track candidate. Some estimators are inherently robust by construction; other estimators can be made robust by finding and discarding outliers.

In the Kalman filter, the residual of the measurement in layer k with respect to the updated state vector is

$$\boldsymbol{r}_{k|k} = \boldsymbol{m}_k - \boldsymbol{h}_k(\boldsymbol{q}_{k|k}),$$

and the covariance matrix of this residual is

$$\boldsymbol{R}_{k|k} = \boldsymbol{V}_k - \boldsymbol{H}_k\boldsymbol{C}_{k|k}\boldsymbol{H}^{\mathrm{T}}.$$

The chi-square increment in layer k is

$$\chi^2_{k,+} = \boldsymbol{r}_{k|k}{}^{\mathrm{T}}\boldsymbol{R}^{-1}_{k|k}\boldsymbol{r}_{k|k},$$

and the total chi-square of the track is found by summing up the chi-square increments for all measurements in the track candidate. The total chi-square is used to evaluate the quality of the track candidate. A too large value of this test statistic indicates that one or more of the measurements of the track candidate do not originate from the particle creating the majority of the measurements. Such measurements are called outliers.

An outlier rejection procedure can make use of the chi-squares of the measurements with respect to the smoothed predictions, i.e., a weighted mean of the predicted states of a forward and a backward Kalman filter. The measurement with the largest value of the chi-square is removed, and the total chi-square is again calculated. This procedure is repeated until the value of the total chi-square falls below a defined threshold.

In the presence of a potentially large fraction of outliers in a track candidate, the sequential outlier rejection procedure outlined above might become unstable, because the smoothed predictions may themselves be biased by outliers. An alternative approach is the Gaussian-sum filter [62]. This algorithm is based on the assumption that the probability distribution of the measurement error can be modeled as a two-component Gaussian mixture, where a narrow component represents the hypothesis that the measurement is real and a wider component represents the hypothesis that the measurement is an outlier. It takes the form of a set of Kalman filters running in parallel, each Kalman filter representing a specific hypothesis of a subset of the measurements that should be classified as outliers. A weight attached to each Kalman filter can be interpreted as the probability of correctness of the hypothesis. In the end, the Kalman filter with the largest weight or a weighted mean of the different filters can be taken as the final estimate.

The Gaussian-sum filter can also be used to deal with a mixture model of the process noise, i.e., the stochastic disturbance of the track because of interactions with the detector material [63]. In the case of bremsstrahlung, a successful application to the reconstruction of electrons is described in [50].

For the treatment of outliers, the Gaussian-sum filter has two disadvantages. First, it may create a large number of Kalman filters running in parallel because of poor knowledge of the track parameters in the early stages of the filter, making the approach expensive in terms of computing time. Second, an explicit outlier model is required. A faster and even more robust alternative is the Deterministic Annealing Filter [64]. This filter is an iterated Kalman filter with annealing, which assigns small weights to measurements far away from the track. A temperature parameter is introduced, facilitating convergence to the globally optimal solution. The iterations start at a high temperature, continue with a gradual lowering of the temperature and converge at the nominal value of the temperature. The procedure is easily generalized to the situation of several measurements being present in the same detector layer. In this case the measurements compete for inclusion in the track. As opposed to a standard outlier rejection approach, the assignment of measurements is soft. This means that several measurements in the same detector layer might contribute to the final estimate of the track parameters, each with a weight equal to the assignment probability. A further generalization is the multi-track filter, where several tracks are allowed to compete for compatible hits in all detector layers [65]. For an experimental application, see [66].

13.1.3.6 Jet Reconstruction

Jets are bundles of collimated hadrons, reflecting hard scattering processes at the parton level. In order to carry out detailed comparisons between parton-level predictions and hadron-level observations a well-defined "jet finder" is required. In the jet finding information from both the tracking devices and the calorimeters is used.

Jet finding can be understood as finding clusters in the set of reconstructed tracks, including neutral tracks. As in the case of vertex finding (see Sect. 13.2.2), various types of clustering methods have been proposed and investigated. The performance strongly depends on the underlying physics, and usually a jet finder is optimized for specific physics requirements. For instance, the widely used $k_\perp$ clustering algorithm comes in several versions, for instance one for e^+e^- collisions [67], and one for hadron-hadron collisions [68].

Hierarchical cluster algorithms offer a large variety of jet finders, differing mainly by the definition of the measure of distance between objects (tracks and jets), but sometimes also by the order in which the objects are combined. Some examples of agglomerative clustering algorithms are described and studied in [69]. Table 13.1 gives a summary of the distance measures used. The names refer to the ones used in [69]. E_i is the energy of cluster i, p_i is its momentum, θ_{ij} is the opening angle between the momentum vectors of the two clusters, and E_{vis} is the visible energy.

A divisive hierarchical clustering algorithm is described in [76]. It is based on the following measure of distance between two tracks:

$$d_{ij} = \frac{\theta_{ij}^2}{p_i p_j},$$

Table 13.1 Some distance measures used for agglomerative jet finding with respective references

Name	Distance d_{ij}	References
Jade	$\dfrac{2\,E_i E_j (1-\cos\theta_{ij})}{E_{\text{vis}}^2}$	[70]
Durham, $k_\perp$	$\dfrac{2\,\min(E_i^2, E_j^2)(1-\cos\theta_{ij})}{E_{\text{vis}}^2}$	[67, 68, 71–73]
Luclus	$\dfrac{2\,p_i^2 p_j^2 (1-\cos\theta_{ij})}{(p_i+p_j)^2 E_{\text{vis}}^2}$	[74]
Geneva	$\dfrac{8}{9}\dfrac{E_i E_j (1-\cos\theta_{ij})}{(E_i+E_j)^2}$	[75]
Cambridge	$\dfrac{2\,\min(E_i^2, E_j^2)(1-\cos\theta_{ij})}{E_{\text{vis}}^2}$	[71]

but can be generalized to any other measure of distance. The method first constructs a minimum spanning tree [77] in the edge-weighted graph connecting all particles with each other and then proceeds to cut the tree along its longest edges. The procedure stops when the longest remaining edge is shorter than a fixed multiple of the median of all edge lengths.

Several non-hierarchical cluster algorithms have been proposed as well. Some of them employ general unsupervised learning methods, such as deterministic annealing [78] or k-means [79]. Others are specially designed for jet finding, for instance the cone algorithm described in [72]. It is an iterating procedure which constructs jets out of seeds. In contrast to the hierarchical clustering method the jets may overlap and a unique assignment has to be forced at the end. A modified cone algorithm suitable for the much larger multiplicity of heavy-ion collisions is proposed in [80]. A specialized jet finder for the reconstruction of hadronic τ-decays is described in [81].

13.1.4 Detector Alignment[1]

Alignment is the general term used in experimental high energy physics to refer to the process of obtaining and applying corrections to the nominal setup of a given experiment. These corrections are typically related to geometrical displacements of devices with a spatial resolution, in contrast to calibrations, where the corrections are usually extracted from pedestal or reference measurements to compensate for offsets in scalar measurements. Misalignment compromises tracking and vertex finding [82] and thus directly affects physics measurements such as momentum and invariant mass resolutions, or the efficiency of b-tagging algorithms. There are various possibilities for the treatment of alignment corrections, ranging from simple translations and rotations, equivalent to those of a rigid body, to more complex deformations, like sags or twists.

To this end experiments typically use several independent strategies [83]. For testing the long-term stability or the alignment of sub-detectors with respect to each other, very often so-called hardware alignment is utilized, where special reference markers are measured directly e.g. via optical systems or photogrammetry. However, these techniques reach only a limited precision in the range of several tens to hundreds of microns. If the intrinsic resolution of a tracking device is smaller, an improved resolution can only be obtained with track-based alignment, where the information from recorded particle tracks is used to obtain the alignment parameters [83, 84]. For various examples of the track-based alignment methods used in experiments since the LEP era, see [85–100].

[1]The section on detector alignment was contributed by E. Widl (Institute of High Energy Physics, Vienna; now at Austrian Institute of Technology).

13.1.4.1 General Overview

The basis of all track-based alignment algorithms is an extended track model $\boldsymbol{d}$, where the measurements $\boldsymbol{m}$ depend not only on the true track-parameters $\boldsymbol{q}_0$, but also on a set of alignment parameters $\boldsymbol{p}_0$ that describe the effects of sufficiently small deviations from the ideal geometry:

$$\boldsymbol{m} = \boldsymbol{d}(\boldsymbol{q}_0, \boldsymbol{p}_0) + \boldsymbol{\gamma}, \quad \mathrm{cov}(\boldsymbol{\gamma}) = \boldsymbol{V}.$$

The stochastic term $\boldsymbol{\gamma}$, which describes the intrinsic resolution of the tracking devices and the effects of multiple scattering, is dealt with via its covariance matrix $\boldsymbol{V}$. Since typically high momentum particles are used, energy-loss effects can be assumed to be deterministic and hence directly taken care of in the track model $\boldsymbol{d}$ itself.

With an initial guess $\breve{\boldsymbol{q}}$ for the track parameters and $\breve{\boldsymbol{p}}$ for the alignment parameters, this model allows to define residuals that are functions of the unknowns $\boldsymbol{q}$ and $\boldsymbol{p}$:

$$\boldsymbol{r}(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{m} - \boldsymbol{d}(\boldsymbol{q}, \boldsymbol{p}) \approx \boldsymbol{m} - \breve{\boldsymbol{d}} - \boldsymbol{D}_q \Delta\boldsymbol{q} - \boldsymbol{D}_p \Delta\boldsymbol{p} \tag{13.5}$$

with

$$\breve{\boldsymbol{d}} = \boldsymbol{d}(\breve{\boldsymbol{q}}, \breve{\boldsymbol{p}}), \quad \Delta\boldsymbol{q} = \boldsymbol{q} - \breve{\boldsymbol{q}}, \quad \Delta\boldsymbol{p} = \boldsymbol{p} - \breve{\boldsymbol{p}},$$

$$\boldsymbol{D}_q = \left.\frac{\partial \boldsymbol{d}}{\partial \boldsymbol{q}}\right|_{\breve{\boldsymbol{q}}, \breve{\boldsymbol{p}}}, \quad \boldsymbol{D}_p = \left.\frac{\partial \boldsymbol{d}}{\partial \boldsymbol{p}}\right|_{\breve{\boldsymbol{q}}, \breve{\boldsymbol{p}}}.$$

The goal of a track-based alignment algorithm is to determine $\boldsymbol{p}$ from the residuals $\boldsymbol{r}$, by minimizing the quadratic form $\chi^2 = \boldsymbol{r}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{r}$, using a sufficiently large set of recorded tracks. The methods used are quite diverse, but can be grouped into two categories: biased and unbiased algorithms.

Biased algorithms initially ignore the fact that the initial guess of the track parameters $\boldsymbol{q}_0$ is in general biased by the factual misalignment. In other words, by setting $\boldsymbol{q} = \breve{\boldsymbol{q}}$ for every track, the residuals become a function of $\boldsymbol{p}$ alone, i.e., $\boldsymbol{r}(\boldsymbol{q}, \boldsymbol{p}) \rightarrow \boldsymbol{r}(\boldsymbol{p})$. In general, the influence of the biased track information has to be compensated by iterating several times over the track sample, where at each iteration step the previously determined parameters are applied to the track reconstruction.

Unbiased algorithms on the other hand, minimize the residuals or the normalized residuals, respectively, estimating at the same time the track parameters. The problem with such an approach is the resulting huge number of parameters. In the presence of N alignment parameters and a sample of M tracks with m track parameters each, a total of $N + m \cdot M$ parameters have to be dealt with. While the value of N depends on the experimental setup, and m usually equals 5, the number of tracks M has always to be of considerable size to acquire reasonable statistics.

On the other hand, unbiased algorithms usually do not require iterations, with the possible exception of problems like non-linearities or rejection of outliers.

Besides the differences between various algorithms it should be noted that the final result of any track-based alignment is always limited by the tracks used. Basic quality cuts, like the selection of high momentum tracks to minimize the influence of multiple scattering or cuts on the minimum number of hits, have a strong influence on the convergence. More subtle is the effect of an unbalanced mixture of tracks or the complete absence of some types of tracks, such as tracks from collisions and cosmic events or tracks taken with and without a magnetic field. This is due to the fact that any kind of tracks has several unconstrained degrees of freedom, usually referred to as weak modes, weakly defined modes or χ^2-invariant modes. As an example, typical weak modes for straight tracks are shears but not bends, and vice versa for curved tracks. Combining the information of both kinds of tracks is therefore a reasonable strategy to avoid these deformations in the final result. The most obvious weak mode is a translation or rotation of the entire tracking device, which can be only fixed with some kind of reference frame, be it an external system or by definition. This, however, is less severe and sometimes even not considered at all, as it does not affect the internal alignment of the tracking device.

Once a set of alignment parameters is calculated, it should always be validated [83, chapter 11]. Apart from checking the improvement of the residuals, several physics measurements can be utilized, especially to probe for remaining weak modes. Known charge, forward-backward or φ-symmetries of distinct physics processes can be used. Distributions of the signed curvature or the signed transverse impact parameter are also sensitive observables.

13.1.4.2 Examples of Alignment Algorithms

Some modern experiments deploy large tracking devices that require a large number of alignment parameters, of the order of 10^5. In such a case the computation of parameters by using straightforward recipes might become unreasonably slow or cause numerical problems. The two algorithms presented in this section are examples of how to cope with such challenging circumstances.

The HIP Algorithm

The HIP algorithm [101] is a straightforward and easy-to-implement biased alignment algorithm. It computes the alignment parameters for each alignable object separately. Only when iterating on the track sample a certain kind of indirect feedback between the alignable objects is established due to the track refit.

Since only individual alignable objects are regarded, Eq. (13.5) can be partitioned. This is simply done by evaluating the corresponding expressions for each alignable object i together with its associated parameters $\boldsymbol{p}_i$:

$$\boldsymbol{r}_i(\boldsymbol{p}_i) = \boldsymbol{m}_i - \boldsymbol{d}_i(\breve{\boldsymbol{q}}, \boldsymbol{p}_i) \approx \boldsymbol{m}_i - \breve{\boldsymbol{d}}_i - \boldsymbol{D}_{pi}\boldsymbol{\Delta p}_i$$

with

$$\breve{\boldsymbol{d}}_i = \boldsymbol{d}_i(\breve{\boldsymbol{q}}, \breve{\boldsymbol{p}}_i), \quad \Delta\boldsymbol{p}_i = \boldsymbol{p}_i - \breve{\boldsymbol{p}}_i, \quad \boldsymbol{D}_{pi} = \left.\frac{\partial \boldsymbol{d}_i}{\partial \boldsymbol{p}_i}\right|_{\breve{\boldsymbol{q}},\, \breve{\boldsymbol{p}}_i}.$$

The result is determined by minimizing the normalized squared residuals from a given set of tracks, again for each alignable object separately. The formal solution is given by

$$\Delta\boldsymbol{p}_i = \left(\sum_{\text{tracks}} \boldsymbol{D}_{pi}{}^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{D}_{pi}\right)^{-1} \left(\sum_{\text{tracks}} \boldsymbol{D}_{pi}{}^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{r}_i(\breve{\boldsymbol{p}}_i)\right)$$

The Millepede Algorithm

The Millepede algorithm [102] is an unbiased algorithm that minimizes the sum of the squared residuals of all tracks at once. To this end a system of linear equations, equivalent to the formal solution of an ordinary χ^2-fit, is solved. However, to achieve this within a reasonable amount of time, only the solution for the alignment parameters is computed, while the computation of the improved track parameters is skipped. This is possible because of the special structure of the system: Firstly, the coefficient matrix is symmetric and, mostly due to the independence of the individual tracks, relatively sparse. Secondly, only the alignment parameters are common parameters for all track measurements, while the specific track parameters are only relevant for each corresponding track. Due to the latter, the solutions for the alignment and track parameters are only coupled via coefficient matrices of the form

$$\boldsymbol{G} = \boldsymbol{D}_p{}^{\mathrm{T}}\, \boldsymbol{V}^{-1} \boldsymbol{D}_q.$$

To set up the reduced system of equations, for each track the following information has to be extracted:

$$\boldsymbol{\Gamma} = \boldsymbol{D}_q{}^{\mathrm{T}}\, \boldsymbol{V}^{-1} \boldsymbol{D}_q, \quad \boldsymbol{\beta} = \boldsymbol{D}_q{}^{\mathrm{T}}\, \boldsymbol{V}^{-1} \left(\boldsymbol{m} - \breve{\boldsymbol{d}} - \boldsymbol{D}_p\, \Delta\boldsymbol{p}'\right).$$

Here $\Delta\boldsymbol{p}' = \boldsymbol{p}' - \breve{\boldsymbol{p}}$ may already include an estimate $\boldsymbol{p}'$ on the actual alignment. Then compute

$$\Delta\boldsymbol{C} = \boldsymbol{D}_p{}^{\mathrm{T}}\, \boldsymbol{V}^{-1} \boldsymbol{D}_p - \boldsymbol{G}\, \boldsymbol{\Gamma}^{-1} \boldsymbol{G}^{\mathrm{T}}, \quad \Delta\boldsymbol{g} = \boldsymbol{D}_p{}^{\mathrm{T}}\, \boldsymbol{V}^{-1} \left(\boldsymbol{m} - \breve{\boldsymbol{d}} - \boldsymbol{D}_p\, \Delta\boldsymbol{p}' + \boldsymbol{D}_q\, \boldsymbol{\Gamma}^{-1} \boldsymbol{\beta}\right).$$

Note the expression $-\boldsymbol{\Gamma}^{-1}\boldsymbol{\beta}$ instead of $\boldsymbol{\Delta q}$. These are all necessary terms, including implicitly the full information from all track parameters. The complete system of equations to determine the alignment parameters then reads

$$\boldsymbol{C}\,\boldsymbol{\Delta p} = -\boldsymbol{g},$$

with

$$\boldsymbol{C} = \sum_{\text{tracks}} \boldsymbol{\Delta C}, \quad \boldsymbol{g} = \sum_{\text{tracks}} \boldsymbol{\Delta g}.$$

The solution by matrix inversion is only feasible if the number of parameters is fairly small ($N \leq 10^3$). The matrix $\boldsymbol{C}$ is usually relatively sparse, so that less time-consuming and more reliable methods can be used, such as the GMRES algorithm [103].

It is also possible to introduce constraints into the solution, which allows to align on various hierarchical levels at once. When aligning for instance on module- and layer-level at the same time, these constraints can remove redundant degrees of freedom by forcing the average movement of all modules within one layer to zero.

Millepede is a well-tested algorithm. To use it efficiently, some knowledge of its inner workings is of advantage. The HIP algorithm is simpler to implement, but less suitable for very large setups than Millepede. Another unbiased algorithm is the Kalman Alignment Algorithm [104]. It is a sequential method, derived from the Kalman filter (see also [105]).

13.1.5 Momentum Resolution

The momentum resolution that can be achieved by a tracking detector is determined by the magnetic field, the arrangement and precision of the tracking detectors, and the amount of material crossed by the particle. Simple approximate formulas can be obtained for two cases:

(a) A spectrometer consisting of a central bending magnet and two arms of tracking detectors in front of and behind the magnet. This is a typical arrangement for a fixed-target experiment with small track multiplicities.
(b) A set of cylindrical tracking detectors immersed in a homogeneous magnetic field. This is a typical arrangement for the barrel part of a collider experiment, for instance layers of silicon or a TPC.

The units are the same as in Sect. 13.1.3.2: momentum in GeV/c, length in meters, and magnetic field in Tesla.

13.1.5.1 Two-arm Spectrometer

We assume that the trajectory of the particle is parallel to the z axis and that B_y is the only significant component of the magnetic field. The angle of deflection is then given by [37]

$$\alpha \approx -\frac{kq}{p}\int_L B_y \,\mathrm{d}z = -\frac{kq}{p}\bar{B}_y L,$$

where L is the length of the magnet, $\bar{B}_y$ is the average value of the field along the trajectory, p is the momentum, and q, k are as in Eq. (13.2). Assuming that $|q| = 1$, linear error propagation gives

$$\frac{\sigma(p)}{p} = \frac{p\sigma(\alpha)}{k|\bar{B}_y|L}.$$

Assume that each arm consists of m identical position detectors spread over a length l, and that the standard deviation of the measurement error of x is equal to δ. The best angular resolution is obtained if in each arm half of the detectors is placed at each end of the arm. Neglecting all multiple scattering, it is equal to

$$\sigma(\alpha) = \frac{2\delta}{l\sqrt{m/2}}.$$

The relative momentum resolution due to measurement errors is therefore

$$\frac{\sigma_{\mathrm{me}}(p)}{p} = \frac{2p\delta}{l\sqrt{m/2}k|\bar{B}_y|L}.$$

Although this arrangement optimizes the precision in terms of geometry, it offers little redundancy for track finding and should be used only in setups with trivial pattern recognition requirements, for instance in the forward direction of fixed target experiments.

At low energies, multiple scattering can no longer be neglected. Whereas $\sigma(p)/p$ arising from position measurement errors only is proportional to p, the term $\sigma_{\mathrm{ms}}(p)/p$ arising from multiple scattering is proportional to $1/(\beta|\bar{B}_y|L)$, which is large for small β and constant for high momenta ($\beta \approx 1$). Under the same assumptions about the detector positions as above, the following formula is obtained:

$$\frac{\sigma_{\mathrm{ms}}(p)}{p} = \frac{0.0136}{\beta k|\bar{B}_y|L}\left(\frac{md}{X_0}\right)^{1/2},$$

where d/X_0 is the thickness of the detectors in units of radiation length. The total resolution is obtained by adding the corresponding variances and taking the square root,

$$\frac{\sigma(p)}{p} = \frac{\sigma_{\text{me}}(p)}{p} \oplus \frac{\sigma_{\text{ms}}(p)}{p} = ap \oplus b,$$

with a and b depending on the detector and the magnetic field.

13.1.5.2 Cylindrical Spectrometer

Assume that there are m cylindrical detectors immersed in a homogeneous magnetic field B_z parallel to the z axis. The projection of the track on the x–y plane is a circle with curvature κ. For high momentum the circle can be approximated by a parabola, the detector cylinders can be approximated by planes, and multiple scattering can be neglected. For this case closed formulas for the joint covariance matrix of κ and the tangent $t_\varphi = \tan\varphi$ of the initial track direction φ can be given [106, 107]. For equidistant detectors, uniform resolution δ and $t_\varphi = 0$ it is given by

$$\text{Cov}\begin{pmatrix} t_\varphi \\ \kappa \end{pmatrix} = \frac{\delta^2}{m(m+1)(m+2)} \begin{pmatrix} \dfrac{12\,(m-1)(2m-1)(8m-11)}{L^2(m-2)} & -\dfrac{360\,(m-1)^3}{L^3(m-2)} \\ -\dfrac{360\,(m-1)^3}{L^3(m-2)} & \dfrac{720\,(m-1)^3}{L^4(m-2)} \end{pmatrix},$$

where L is now the track length in the x–y projection. L is approximately equal to the radial distance between the innermost and the outermost detector. As $\kappa = kB_z/p_{\text{T}}$,

$$\frac{\sigma_{\text{me}}(p_{\text{T}})}{p_{\text{T}}} = \frac{p_{\text{T}}}{k|B_z|L}\frac{\delta}{L}\left[\frac{720\,(m-1)^3}{(m-2)m(m+1)(m+2)}\right]^{1/2}.$$

There is a high negative correlation between $1/p_{\text{T}}$ and the direction tangent t_φ. For large m, the asymptotic values are

$$\frac{\sigma_{\text{me}}(p_{\text{T}})}{p_{\text{T}}} = \frac{p_{\text{T}}}{k|B_z|L}\frac{\delta}{L}\left[\frac{720}{m+4}\right]^{1/2}, \quad \sigma(t_\varphi) = \frac{\delta}{L}\left[\frac{192}{m+3.875}\right]^{1/2},$$

$$\text{cov}(t_\varphi, 1/p_{\text{T}}) = -\frac{\sqrt{15}}{4} = -0.968.$$

More general closed formulas for $t_\varphi \neq 0$ are given in [107].

Table 13.2 Values of C_m in Eq. (13.6)

m	3	4	5	6	7	8	9	10	> 10 m
C_m	1.16	1.06	1.04	1.03	1.02	1.02	1.01	1.01	≈ 1 m

If one half of the detectors is placed at the center of the track and one quarter at either end, the variance of the curvature is minimal, and the covariance matrix reads

$$\mathrm{Cov}\begin{pmatrix} t_\varphi \\ \kappa \end{pmatrix} = \delta^2 \begin{pmatrix} \dfrac{72}{L^2 m} & -\dfrac{128}{L^3 m} \\ -\dfrac{128}{L^3 m} & \dfrac{256}{L^4 m} \end{pmatrix},$$

which is considerably smaller than in the equidistant case. This arrangement, however, is not particularly well suited for track finding and moreover difficult to realize.

The contribution of multiple scattering to the transverse momentum resolution can be approximated by

$$\frac{\sigma_{\mathrm{ms}}(p_{\mathrm{T}})}{p_{\mathrm{T}}} = C_m \cdot \frac{s}{\beta k |B_z| L} \left(\frac{md}{X_0 \cos\lambda} \right)^{1/2}, \tag{13.6}$$

where d/X_0 is the thickness of the detectors in units of radiation length, $\lambda = \pi/2 - \theta$ is the dip angle of the track, $s = 0.0136(1 + 0.038 \ln(d/X_0))$, k is as in Eq. (13.2), and C_m is a factor depending on m. Values of C_m for small m, obtained by the program described in [108], are given in Table 13.2. Note that the values are different from the ones given in [106]. In a time projection chamber md/X_0 has to be replaced by L/X_0, where X_0 is the radiation length of the gas. The factor $\cos\lambda$ in the denominator accounts for the actual amount of matter traversed by a track with dip angle λ. Approximate formulas for the best possible resolution including multiple scattering can be found in [109].

The total transverse momentum resolution is calculated by quadratic addition,

$$\frac{\sigma(p_{\mathrm{T}})}{p_{\mathrm{T}}} = \frac{\sigma_{\mathrm{me}}(p_{\mathrm{T}})}{p_{\mathrm{T}}} \oplus \frac{\sigma_{\mathrm{ms}}(p_{\mathrm{T}})}{p_{\mathrm{T}}},$$

which can be written in the form

$$\frac{\sigma(p_{\mathrm{T}})}{p_{\mathrm{T}}} = \frac{a\, p_{\mathrm{T}}}{\sqrt{m+4}} \oplus \frac{b\sqrt{m}}{\sqrt{\cos\lambda}}.$$

This shows that an optimal m exists for every p_{T} and λ if the projected track length L is kept fixed. Overinstrumentation will deteriorate the resolution for low momenta unless additional measurements can be included without increasing the amount of matter to be traversed.

In order to calculate the error of the momentum $p = p_T/\cos\lambda$ the error in λ must be taken into account:

$$\sigma^2(p) = \sigma^2(p_T)/\cos^2\lambda + \sigma^2(\lambda)\, {p_T}^2 \sin^2\lambda/\cos^4\lambda,$$

the correlation between p_T and λ being negligible in practice. Because of $\sigma(p)/p = p\sigma(1/p)$ it follows that:

$$\frac{\sigma(p)}{p} = \frac{\sigma(p_T)}{p_T} \oplus \sigma(\lambda)\tan\lambda.$$

With the exception of very low momenta the track can be approximated by a straight line in the r–z projection, where $r = (x^2 + y^2)^{1/2}$. For m equidistant detectors and uniform resolution δ, the variance of the direction tangent $t_\lambda = \tan\lambda$ due to the measurement errors is given by [106]:

$$\sigma^2_{\text{me}}(t_\lambda) = \frac{\delta^2}{L^2}\frac{12(m-1)}{m(m+1)}\frac{1}{\cos^4\lambda}.$$

If the measurement error in z is very small, the variance of t_λ is dominated by multiple scattering. For equidistant layers of uniform thickness d an approximate formula can be given. Remarkably, it does not depend on the number of layers:

$$\sigma^2_{\text{ms}}(t_\lambda) \approx \frac{s^2}{p^2}\frac{d}{X_0\cos\lambda}\frac{1}{\cos^4\lambda} = \frac{s^2}{{p_T}^2}\frac{d}{X_0\cos\lambda}\frac{1}{\cos^2\lambda},$$

with $s = 0.0136(1 + 0.038\ln(d/X_0))$.

In the design and optimization phase of the detector a precise evaluation of the resolution of all track parameters is mandatory. There are several software packages that allow a fast track simulation plus reconstruction in a general detector setup, for instance [110] (in FORTRAN), [111] (in Matlab/Octave), or [108] (in Java).

13.2 Vertex Reconstruction

13.2.1 Introduction

Vertex reconstruction is the task of finding and estimating the production point of a set of particles. The pattern recognition algorithms and statistical estimation methods involved are in many respects similar to the ones used in track reconstruction. For an overview of vertex reconstruction algorithms used in past or active experiments see for instance [112–116].

In practice it is useful to distinguish between several types of vertices:

1. The primary vertex is the point of collision of two beam particles (in a collider experiment) or of a beam particle and a target particle (in a fixed-target experiment).
2. A secondary decay vertex is the point where an unstable particle decays in the detector volume or in the beam pipe. An example is the decay $K_S^0 \to \pi^+\pi^-$.
3. A secondary interaction vertex is the point where a particle interacts with the material of the detector. Examples are bremsstrahlung, pair production, and inelastic hadronic interactions.

Vertex reconstruction frequently proceeds in several steps:

1. Vertex finding: Finds the tracks that belong to a common primary or secondary vertex.
2. Vertex fitting: Estimates for each vertex candidate the location of the common vertex and computes the associated covariance matrix.
3. Test of vertex hypothesis: Tests for each vertex candidate whether all tracks do indeed belong to the vertex and identifies outliers.
4. Update: Uses the vertex constraint to improve the location and momentum estimate of the tracks belonging to the vertex.
5. Kinematic fit: Kinematic constraints such as momentum and energy conservation are imposed on the mother and daughter particles of a vertex, and mass hypotheses are tested. Kinematic fits are most frequently applied to secondary decay vertices.

Vertex finding can be accomplished in many different ways. A few of them will be described in Sect. 13.2.2. The vertex fit takes a vertex candidate and estimates the vertex location from the estimated track parameters of the outgoing particles (Sect. 13.2.3). As a rule, only charged particles are used, but sometimes also neutral particles contribute to the vertex fit. In the test stage (Sect. 13.2.3.2) outliers are identified, i.e., particles that apparently do not belong to the estimated vertex. As this can lead to a different assignment of particles to vertices, it can be considered as a method of vertex finding. If outliers are expected, the estimation procedure should be robust so that the estimated vertex is not significantly biased by the outliers (Sect. 13.2.3.3). Kinematic constraints (Sect. 13.2.4) are usually imposed via Lagrange multipliers. By repeating the kinematic fit under various mass hypotheses of the mother and/or daughter particles the most likely mass assignment can be found out.

13.2.2 Vertex Finding

Vertex finding is the process of dividing the reconstructed tracks in an event into classes such that presumably all tracks in a class are produced at the same vertex. The primary vertex in an event is usually easy to find, especially if prior information

about its location is available (beam profile, target position). On the other hand, secondary decay vertices of short-lived decays are hard to find, as some of the decay products may also be compatible with the primary vertex. Vertex finding methods can be roughly divided in three main types: generic clustering algorithms, topological methods, and iterated estimators. The latter can be considered as a special divisive clustering method.

13.2.2.1 Clustering Methods

As mentioned above in the context of jet finding (see Sect. 13.1.3.6), clustering methods are based on a distance matrix or a similarity matrix of the objects to be classified. A cluster is then a group with small distances (large similarities) inside the group and large distances (small similarities) to objects outside the group. The distance measure reflects only the geometry of the tracks.

Various clustering methods have been evaluated in the context of vertex finding, of both the hierarchical and the non-hierarchical type [117]. Hierarchical clustering can be agglomerative or divisive. In agglomerative clustering each track starts out as a single cluster. Clusters are merged iteratively on the basis of a distance measure. The shortest distance in space between two tracks is peculiar insofar as it does not satisfy the triangle inequality: if tracks a and b are close, and tracks b and c are close, it does not follow that tracks a and c are close as well. The distance between two clusters of tracks should therefore be defined as the maximum of the individual pairwise distances, known as complete linkage in the clustering literature. Alternatively, the distance between two clusters can be the distance between the two vertices fitted from the clusters. Divisive clustering starts out with a single cluster containing all tracks. Further division of this cluster can be based on repeated vertex estimation with outlier identification (see Sect. 13.2.2.3). Examples of non-hierarchical clustering methods used in vertex finding are vector quantization, the k-means algorithm and deterministic annealing [113].

13.2.2.2 Topological Methods

A very general topological vertex finder was proposed in [118]. It is related to the Radon transform, which is a continuous version of the Hough transform used for track finding (Sect. 13.1.2.1). The search for vertices is based on a function $V(\boldsymbol{v})$ which quantifies the probability of a vertex at location $\boldsymbol{v}$. For each track a Gaussian probability tube $f_i(\boldsymbol{v})$ is constructed. The function $V(\boldsymbol{v})$ is defined taking into account that the value of $f_i(\boldsymbol{v})$ must be significant for at least two tracks:

$$V(\boldsymbol{v}) = \sum_{i=0}^{n} f_i(\boldsymbol{v}) - \frac{\sum_{i=0}^{n} f_i^2(\boldsymbol{v})}{\sum_{i=0}^{n} f_i(\boldsymbol{v})}$$

Due to the second term on the right-hand side, $V(\boldsymbol{v}) \approx 0$ in regions where $f_i(\boldsymbol{v})$ is significant for only one track. The form of $V(\boldsymbol{v})$ can be modified to fold in known physics information about probable vertex locations. For instance, $V(\boldsymbol{v})$ can be augmented by a further function $f_0(\boldsymbol{v})$ describing the location and spread of the interaction point. In addition, $V(\boldsymbol{v})$ may be modified by a factor dependent on the angular location of the point $\boldsymbol{v}$.

Vertex finding amounts to finding local maxima of the function $V(\boldsymbol{v})$. The search starts at the calculated maxima of the products $f_i(\boldsymbol{v})f_j(\boldsymbol{v})$ for all track pairs. For each of these points the nearest maximum of $V(\boldsymbol{v})$ is found. These maxima are clustered together to form candidate vertex regions. The final association of the tracks to the vertex candidates can be done on the basis of the respective χ^2 contributions or by an adaptive fit (see Sect. 13.2.3.3). In [119] the topological vertex finder was augmented by a procedure based on the concept of the minimum spanning tree of a graph.

13.2.2.3 Iterated Estimators

Vertex finding can also be accomplished by iterated vertex fits (see Sect. 13.2.3). The procedure can be summarized in the following way:

1. Fit one vertex with all tracks
2. Discard all incompatible tracks
3. Repeat step 1 with all discarded tracks

The iteration stops when no vertex with at least two tracks can be successfully fitted. Step 2 might itself be iterative, especially if the vertex fit is not robust, so that the incompatible tracks have to be removed sequentially. Iterative vertex finders based on a least-squares fit (Sect. 13.2.3.1) and an adaptive fit (Sect. 13.2.3.3) are implemented in the RAVE toolbox [120, 121].

13.2.3 Vertex Fitting

The input to the vertex fit is a vertex candidate, i.e., a set of estimated track parameters $\{\tilde{\boldsymbol{q}}_1, \ldots, \tilde{\boldsymbol{q}}_n\}$ located at one or more reference surfaces, along with their covariance matrices $\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n\}$. For instance, in the primary vertex fit in a collider experiment the reference surface may be the beam tube. If possible, the reference surface(s) should be chosen such that multiple scattering between the vertex and the location of the track parameters is negligible.

The parameters to be fitted are the vertex position $\boldsymbol{v}$ and the track momenta $\boldsymbol{p}_i$ at the vertex. The functional dependence of the track parameters on the vertex parameters requires a track model, which depends on the shape of the magnetic field in the vicinity of the vertex. If the field is homogeneous, the track model is a helix;

if the field is zero, the track model is a straight line. In other cases the track model may have to be computed numerically (see Sect. 13.1.3.2).

13.2.3.1 Least-Squares Methods

The conventional approach to estimating the vertex position is the minimization of some quadratic objective function, yielding a least-squares estimate. There are two main flavors of least-squares estimation in vertex fitting, constrained and unconstrained minimization. In the first case the vertex constraint is introduced into the objective function via a Lagrange multiplier, in the second case the constraint is implicit in the track model.

As an example, consider a vertex fit with n straight tracks. The n straight tracks originating from the common vertex $\boldsymbol{v} = (x_\mathrm{v}, y_\mathrm{v}, z_\mathrm{v})^\mathrm{T}$ can be represented by n straight lines with parameters λ_i:

$$x = x_\mathrm{v} + \lambda_i a_i, \quad y = y_\mathrm{v} + \lambda_i b_i, \quad z = z_\mathrm{v} + \lambda_i, \quad i = 1, \ldots, n,$$

where a_i and b_i are the direction tangents at the vertex. At the reference surface $z = z_\mathrm{ref}$ track i is specified by its parameter vector $\boldsymbol{q}_i = (x_i, y_i, a_i, b_i)^\mathrm{T}$, consisting of the intersection point (x_i, y_i) and the two direction tangents (a_i, b_i). The track fit delivers estimates $\tilde{\boldsymbol{q}}_i$ and information matrices $\boldsymbol{G}_i$ for $i = 1, \ldots, n$. In the constrained problem the sum of the squared residuals

$$M(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n) = \sum_{i=1}^{n} \boldsymbol{e}_i{}^\mathrm{T} \boldsymbol{G}_i \boldsymbol{e}_i, \quad \boldsymbol{e}_i = \tilde{\boldsymbol{q}}_i - \boldsymbol{q}_i, \tag{13.7}$$

must be minimized under the $2n$ nonlinear constraints

$$x_\mathrm{v} = x_i + a_i(z - z_\mathrm{ref}), \quad y_\mathrm{v} = y_i + b_i(z - z_\mathrm{ref}), \quad i = 1, \ldots, n$$

There are $4n + 3$ unknowns, $4n$ observations and $2n$ constraints, giving $4n + 2n - (4n + 3) = 2n - 3$ degrees of freedom. The resulting track parameters $\bar{\boldsymbol{q}}_i$ fit best, in the least-squares sense, to the track fit estimates $\tilde{\boldsymbol{q}}_i$ and at the same time have a common vertex. For the solution of the constrained vertex fit see Sect. 13.2.4.2.

In the example, the constraints can be rewritten as

$$x_i = x_\mathrm{v} + (z_\mathrm{ref} - z)a_i, \quad y_i = y_\mathrm{v} + (z_\mathrm{ref} - z)b_i, \quad i = 1, \ldots, n. \tag{13.8}$$

Insertion of Eq. (13.8) into Eq. (13.7) gives the objective function of the unconstrained nonlinear least-squares problem:

$$M(\boldsymbol{v}, a_1, b_1, \ldots, a_n, b_n) = \sum_{i=1}^{n} \boldsymbol{e}_i{}^\mathrm{T} \boldsymbol{G}_i \boldsymbol{e}_i, \quad \boldsymbol{e}_i = \tilde{\boldsymbol{q}}_i - \boldsymbol{q}_i.$$

There are now $4n$ observations and $2n + 3$ unknown parameters, namely the vertex position and the track directions at the vertex, giving again $4n - (2n + 3) = 2n - 3$ degrees of freedom.

A generalization of this simple case to helix tracks can be found in [122–124]. In the general case, the unconstrained problem can be formulated in terms of the unknown vertex position $\boldsymbol{v}$ and the unknown track momentum vectors $\boldsymbol{p}_i$ at the vertex [26, 59]. The measurement equation reads

$$\boldsymbol{q}_i = \boldsymbol{h}_i(\boldsymbol{v}, \boldsymbol{p}_i), i = 1, \ldots, n, \tag{13.9}$$

where the function $\boldsymbol{h}_i$ incorporates the track model in the magnetic field. The objective function is equal to

$$M(\boldsymbol{v}, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n) = \sum_{i=1}^{n} \boldsymbol{e}_i{}^{\mathrm{T}} \boldsymbol{G}_i \boldsymbol{e}_i, \quad \boldsymbol{e}_i = \tilde{\boldsymbol{q}}_i - \boldsymbol{q}_i.$$

Minimization of the objective function can proceed in several ways. For a detailed exposition of non-linear least-squares estimation see e.g. [125].

Gauss-Newton Method

Assume that there are approximate values $\breve{\boldsymbol{v}}$ and $\breve{\boldsymbol{p}}_i$ for all i. Then Eq. (13.9) can be approximated by an affine function:

$$\boldsymbol{q}_i \approx \boldsymbol{h}_i(\breve{\boldsymbol{v}}, \breve{\boldsymbol{p}}_i) + \boldsymbol{A}_i(\boldsymbol{v} - \breve{\boldsymbol{v}}) + \boldsymbol{B}_i(\boldsymbol{p}_i - \breve{\boldsymbol{p}}_i) = \boldsymbol{c}_i + \boldsymbol{A}_i \boldsymbol{v} + \boldsymbol{B}_i \boldsymbol{p}_i,$$

with

$$\boldsymbol{A}_i = \left.\frac{\partial \boldsymbol{h}_i(\boldsymbol{v}, \boldsymbol{p}_i)}{\partial \boldsymbol{v}}\right|_{\breve{\boldsymbol{v}}, \breve{\boldsymbol{p}}_i}, \quad \boldsymbol{B}_i = \left.\frac{\partial \boldsymbol{h}_i(\boldsymbol{v}, \boldsymbol{p}_i)}{\partial \boldsymbol{p}_i}\right|_{\breve{\boldsymbol{v}}, \breve{\boldsymbol{p}}_i}, \quad \boldsymbol{c}_i = \boldsymbol{h}_i(\breve{\boldsymbol{v}}, \breve{\boldsymbol{p}}_i) - \boldsymbol{A}_i \breve{\boldsymbol{v}} - \boldsymbol{B}_i \breve{\boldsymbol{p}}_i.$$

The objective function then reads

$$M(\boldsymbol{v}, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n) = \sum_{i=1}^{n} (\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i - \boldsymbol{A}_i \boldsymbol{v} - \boldsymbol{B}_i \boldsymbol{p}_i)^{\mathrm{T}} \boldsymbol{G}_i (\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i - \boldsymbol{A}_i \boldsymbol{v} - \boldsymbol{B}_i \boldsymbol{p}_i).$$

As M is now quadratic in the unknown parameters, the minimum can be computed explicitly. The estimated vertex position and its covariance matrix are given by

$$\tilde{\boldsymbol{v}}_n = \boldsymbol{C}_n \sum_{i=1}^{n} \boldsymbol{A}_i{}^{\mathrm{T}} \boldsymbol{G}_i{}^{B} (\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i), \quad \mathrm{Var}(\tilde{\boldsymbol{v}}_n) = \boldsymbol{C}_n = \left(\sum_{i=1}^{n} \boldsymbol{A}_i{}^{\mathrm{T}} \boldsymbol{G}_i{}^{B} \boldsymbol{A}_i \right)^{-1}, \tag{13.10}$$

with

$$\boldsymbol{G}_i{}^B = \boldsymbol{G}_i - \boldsymbol{G}_i\boldsymbol{B}_i\boldsymbol{W}_i\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i, \quad \boldsymbol{W}_i = (\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{B}_i)^{-1}.$$

In general, the procedure has to be iterated. The measurement equation is expanded at the new estimate, and the estimate is recomputed until convergence is obtained. The formulas required for the implementation of two important cases, fixed-target configuration and solenoidal configuration, are given in [61].

Once $\tilde{\boldsymbol{v}}_n$ is known, the track momenta and the full covariance matrix can be computed:

$$\begin{aligned}
&\tilde{\boldsymbol{p}}_i^n = \boldsymbol{W}_i\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i(\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i - \boldsymbol{A}_i\tilde{\boldsymbol{v}}_n),\\
&\mathrm{Var}(\tilde{\boldsymbol{p}}_i^n) = \boldsymbol{D}_i^n = \boldsymbol{W}_i + \boldsymbol{W}_i\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{A}_i\boldsymbol{C}_n\boldsymbol{A}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{B}_i\boldsymbol{W}_i,\\
&\mathrm{Cov}(\tilde{\boldsymbol{p}}_i^n, \tilde{\boldsymbol{v}}_n) = \boldsymbol{E}_i^n = -\boldsymbol{W}_i\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{A}_i\boldsymbol{C}_n.
\end{aligned} \tag{13.11}$$

The estimates can also be computed recursively, resulting in an extended Kalman filter [25, 26, 59]:

$$\begin{aligned}
&\tilde{\boldsymbol{v}}_i = \boldsymbol{C}_i[\boldsymbol{C}_{i-1}^{-1}\tilde{\boldsymbol{v}}_{i-1} + \boldsymbol{A}_i{}^{\mathrm{T}}\boldsymbol{G}_i{}^B(\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i)], \quad && \boldsymbol{C}_i = (\boldsymbol{C}_{i-1}^{-1} + \boldsymbol{A}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{A}_i)^{-1}\\
&\tilde{\boldsymbol{p}}_i = \boldsymbol{W}_i\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i(\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i - \boldsymbol{A}_i\tilde{\boldsymbol{v}}_i), && \boldsymbol{D}_i = \boldsymbol{W}_i + \boldsymbol{W}_i\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{A}_i\boldsymbol{C}_i\boldsymbol{A}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{B}_i\boldsymbol{W}_i,\\
&\boldsymbol{E}_i = -\boldsymbol{W}_i\boldsymbol{B}_i{}^{\mathrm{T}}\boldsymbol{G}_i\boldsymbol{A}_i\boldsymbol{C}_i.
\end{aligned}$$

The associated smoother is tantamount to recomputing the track momenta using the last vertex estimate $\tilde{\boldsymbol{v}}_n$, i.e., Eq. (13.11).

Newton–Raphson Method

This method uses a local quadratic approximation to the objective function. In order to simplify the notation we introduce $\boldsymbol{\alpha} = (\boldsymbol{v}, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)^{\mathrm{T}}$, $\tilde{\boldsymbol{q}} = (\tilde{\boldsymbol{q}}_1, \ldots, \tilde{\boldsymbol{q}}_n)^{\mathrm{T}}$ and $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n)^{\mathrm{T}}$. Then the objective function can be written as

$$M(\boldsymbol{\alpha}) = [\tilde{\boldsymbol{q}} - \boldsymbol{h}(\boldsymbol{\alpha})]^{\mathrm{T}}\boldsymbol{G}[\tilde{\boldsymbol{q}} - \boldsymbol{h}(\boldsymbol{\alpha})], \quad \boldsymbol{G} = \mathrm{diag}(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_n).$$

If $\breve{\boldsymbol{\alpha}}$ is an appropriate expansion point, $M(\boldsymbol{\alpha})$ is approximated by

$$M(\boldsymbol{\alpha}) \approx M(\breve{\boldsymbol{\alpha}}) + \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{\alpha} - \breve{\boldsymbol{\alpha}}) + \tfrac{1}{2}(\boldsymbol{\alpha} - \breve{\boldsymbol{\alpha}})^{\mathrm{T}}\boldsymbol{\Omega}(\boldsymbol{\alpha} - \breve{\boldsymbol{\alpha}}),$$

where

$$\boldsymbol{g} = \frac{\partial M}{\partial \boldsymbol{\alpha}} = -2\boldsymbol{H}^{\mathrm{T}}\boldsymbol{G}[\tilde{\boldsymbol{q}} - \boldsymbol{h}(\breve{\boldsymbol{\alpha}})], \quad \boldsymbol{\Omega} = \frac{\partial^2 M}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^{\mathrm{T}}} = 2\boldsymbol{H}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{H} - 2\frac{\partial \boldsymbol{H}^{\mathrm{T}}}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}\boldsymbol{G}[\tilde{\boldsymbol{q}} - \boldsymbol{h}(\breve{\boldsymbol{\alpha}})]$$

are the gradient and the Hessian of M, respectively, evaluated at $\breve{\boldsymbol{\alpha}}$, and $\boldsymbol{H}$ is the Jacobian of the track model $\boldsymbol{h}(\boldsymbol{\alpha})$. If $\boldsymbol{\Omega}$ is positive definite, M has a minimum when its gradient is zero, leading to

$$\tilde{\boldsymbol{\alpha}} = \breve{\boldsymbol{\alpha}} - \boldsymbol{\Omega}^{-1}\boldsymbol{g}.$$

If the second term of the Hessian is set to zero, the Gauss–Newton method is recovered. Clearly, the Newton–Raphson method is more complex, but it gives some additional information about the problem. In particular, a Hessian that is not positive definite indicates that the expansion point is too far from the true global minimum.

Levenberg–Marquardt Method

In this method the matrix $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{H}$ is inflated by a diagonal matrix $k\boldsymbol{I}$. As a consequence, the direction of the parameter update is intermediate between the direction of the Gauss–Newton step ($k = 0$) and the direction of steepest descent ($k \to \infty$). An example of a vertex fit with the Levenberg–Marquardt method is given in [122].

Fast Vertex Fits

The estimated track parameters $\tilde{\boldsymbol{q}}_i$ are frequently given at the innermost detector surface or at the beam tube. If the $\tilde{\boldsymbol{q}}_i$ are propagated to the vicinity of the presumed vertex, the vertex estimation can be speeded up by applying some approximations.

The "perigee" parametrization for helical tracks was introduced in [123], with a correction in [124]. The track is parameterized around the point of closest approach (the perigee point $\boldsymbol{v}^{\mathrm{P}}$) of the helix to the z-axis. The variation of transverse errors along the track is neglected in the vicinity of the perigee, and the track direction and curvature at the vertex is considered to be constant. The approximate objective function of the vertex fit can then be written entirely in terms of the perigee points:

$$M(\boldsymbol{v}) = \sum_{i=1}^{n} (\boldsymbol{v}_i{}^{\mathrm{P}} - \boldsymbol{v})^{\mathrm{T}} \boldsymbol{T}_i (\boldsymbol{v}_i{}^{\mathrm{P}} - \boldsymbol{v}), \tag{13.12}$$

where $\boldsymbol{T}_i$ is a weight matrix of rank 2. The vertex estimate is then

$$\tilde{\boldsymbol{v}} = \left(\sum_{i=1}^{n} \boldsymbol{T}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{T}_i \boldsymbol{v}_i{}^{\mathrm{P}} \right).$$

The Jacobians required to compute the $\boldsymbol{T}_i$ are spelled out in [123, 124].

A further simplification was proposed in [126]. In the vicinity of the vertex the track is approximated by a straight line. The estimated track parameters are transformed to a coordinate system the x-axis of which is parallel to the track. The vertex is then estimated by minimizing the sum of the weighted transverse distances of the tracks to the vertex. The resulting objective function has the same form as in Eq. (13.12), again with weight matrices of rank 2. The estimate is exact for straight tracks.

A different type of a fast vertex fitting algorithm is described in [127]. It is based on approximating the tracks by straight lines both in the x–y plane and in the x–z plane. In either projection, the lines representing the tracks are Hough-transformed to points in the dual plane of line parameters. The vertex coordinates are then obtained by a weighted linear least-squares fit in the dual plane.

Adding Prior Information

If the vertex to be fitted is the primary vertex, there may be prior information about the vertex position from the beam profile in a collider experiment or the target location in a fixed target experiment. The prior information usually comes in the form of a position $\boldsymbol{v}_0$ plus a covariance matrix $\boldsymbol{C}_0$. The objective function is then augmented by an additional term

$$(\boldsymbol{v}_0 - \boldsymbol{v})^{\mathrm{T}} \boldsymbol{C}_0^{-1} (\boldsymbol{v}_0 - \boldsymbol{v}).$$

For instance, the Gauss–Newton estimate Eq. (13.10) is modified in the following way:

$$\tilde{\boldsymbol{v}}_n = \boldsymbol{C}_n \left[C_0^{-1} \boldsymbol{v}_0 + \sum_{i=1}^{n} \boldsymbol{A}_i{}^{\mathrm{T}} \boldsymbol{G}_i{}^{B} (\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i) \right], \quad \mathrm{Var}(\tilde{\boldsymbol{v}}_n) = \boldsymbol{C}_n = \left(C_0^{-1} + \sum_{i=1}^{n} \boldsymbol{A}_i{}^{\mathrm{T}} \boldsymbol{G}_i{}^{B} \boldsymbol{A}_i \right)^{-1}.$$

Similar modifications apply to the Newton–Raphson estimate and the fast vertex fits.

13.2.3.2 Vertex Quality and Outlier Removal

Some tracks used in the vertex fit may be outliers in the sense that they do not actually belong to the vertex. Also, the estimated track parameters may be distorted by outliers or distorted hits in the track fit. Both types of outliers distort the vertex estimate and need to be identified.

In the case of Gaussian errors and a linear model the contribution of each track to the minimum value of the objective function is distributed according to a χ^2-distribution with two degrees of freedom. The contribution χ_i^2 of track i has to be computed relative to the vertex estimated without track i. For instance, in the

Gauss–Newton algorithm:

$$\chi_i^2 = \boldsymbol{r}_i^{n\mathrm{T}} \boldsymbol{G}_i \boldsymbol{r}_i^n + (\tilde{\boldsymbol{v}}_n - \tilde{\boldsymbol{v}}_n^{-i})^{\mathrm{T}} (\boldsymbol{C}_n^{-i})^{-1} (\tilde{\boldsymbol{v}}_n - \tilde{\boldsymbol{v}}_n^{-i}),$$

where $\boldsymbol{r}_i^n = \tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i - \boldsymbol{A}_i \tilde{\boldsymbol{v}}_n - \boldsymbol{B}_i \tilde{\boldsymbol{p}}_i^n$ is the residual of track i and $\tilde{\boldsymbol{v}}_n^{-i}$ is the vertex estimate with track i removed:

$$\tilde{\boldsymbol{v}}_n^{-i} = \boldsymbol{C}_n^{-i} \left[\boldsymbol{C}_n^{-1} \tilde{\boldsymbol{v}}_n - \boldsymbol{A}_i{}^{\mathrm{T}} \boldsymbol{G}_i{}^{B} (\tilde{\boldsymbol{q}}_i - \boldsymbol{c}_i) \right], \quad \boldsymbol{C}_n^{-i} = \left(\boldsymbol{C}_n^{-1} - \boldsymbol{A}_i{}^{\mathrm{T}} \boldsymbol{G}_i{}^{B} \boldsymbol{A}_i \right)^{-1}.$$

Analogous but somewhat simpler formulas hold for the fast vertex fits.

The test statistic χ_i^2 can be computed for all i, and the track with the largest χ_i^2 is a candidate for removal. This procedure can be repeated until all χ_i^2 are below the cut. Even if there is only a single outlier, all χ_i^2 are no longer χ^2-distributed and the power of the test is impaired. This loss of power can be compensated by robust estimation of the vertex.

13.2.3.3 Robust and Adaptive Estimators

Robust estimators are less influenced or not influenced at all by outlying observations. This can be achieved by downweighting outliers or by excluding them from the estimate. For example, in the case of a one-dimensional location estimate, the M-estimator [128] downweights outliers, whereas the LMS (least median of squares) estimator [129] uses only one half of the sample (the one spanning the shortest interval) and ignores the other one.

Robust estimators tend to be statistically less efficient and computationally more expensive than least-squares estimators. On the other hand, estimation and outlier detection are performed in parallel, whereas a least-squares estimator has to be recomputed after an outlier has been identified and removed.

One of the earliest proposals for a robust vertex fit is in [130]. The method is an M-estimator with Huber's ψ-function [131]. It is implemented as a re-weighted least-squares estimator. The initial vertex estimate is a plain least-squares estimate. Then, for each track, the residuals are rotated to the eigensystem of the covariance matrix of the track, and weight factors are computed according to

$$w_i = \frac{\psi(r_i/\sigma_i)}{r_i/\sigma_i} = \begin{cases} 1, & |r_i| \le c\sigma_i, \\ c\sigma_i/|r_i|, & |r_i| > c\sigma_i, \end{cases}$$

where r_i is one of the residuals in the rotated frame, σ_i is the standard deviation in the rotated frame, and c is the robustness constant, usually chosen between 1 and 3. The weight factors are applied and the estimate is recomputed. The entire procedure is iterated until convergence.

A different kind of re-weighted least-squares estimator is proposed in [132]. The weights are computed according to Tukey's bi-square function [128]:

$$w_i = \begin{cases} \left(1 - \dfrac{r_i^2/\sigma_i^2}{c^2}\right)^2, & |r_i| \le c\sigma_i, \\ 0, & \text{otherwise}, \end{cases}$$

where r_i^2 is the squared residual of track i with respect to the vertex, σ_i^2 is its variance, and c is again the robustness constant. The estimator is now equivalent to a redescending M-estimator, and consequently less sensitive to outliers than Huber's M-estimator.

The combination of a redescending M-estimator with the concept of deterministic annealing [133] leads to the adaptive method of vertex fitting [113, 117, 134, 135]. The concept of the adaptive vertex fit is derived from the Deterministic Annealing Filter [64] (see Sect. 13.1.3.5). The weights are computed according to

$$w_i = \frac{\exp\left(-\chi_i^2/2T\right)}{\exp\left(-\chi_i^2/2T\right) + \exp\left(-\chi_{\text{cut}}^2/2T\right)},$$

where χ_i^2 is the χ^2-contribution of track i, χ_{cut}^2 is a cutoff value, and T is a temperature parameter. The computation of the redescending M-estimator can be interpreted as an EM (expectation–maximization) algorithm [136, 137]. Alternatively it can be viewed as the minimization of the energy function of an elastic arm algorithm [18, 19]. If annealing is employed, the iteration starts at high T. The temperature is then gradually decreased. At low T the weights approach either zero or one. The final weights can be used for classification of the tracks as inliers or outliers. A comparison of the adaptive method with other robust estimators can be found in [138]. The adaptive estimator has been extended to a multi-vertex estimator fitting several vertices simultaneously, including competition of all vertices for all tracks [139].

Iterated re-weighted least-squares estimators require a good starting point, in order to ensure convergence to the global minimum and to minimize the number of iterations required. In many cases a standard least-squares estimate is sufficient. In the presence of a large number of outliers also the starting point should be estimated robustly, preferably by an estimator with a high breakdown point [129]. Several such initial estimators have been proposed and studied in [117].

The M-estimators and the adaptive estimator presented above do not presuppose an explicit outlier model. If it is possible to describe the outliers by a Gaussian mixture model, estimation of the vertex can be carried out by the Gaussian-sum filter [140].

Several of the estimators described here are implemented in RAVE, a detector-independent toolkit for reconstruction of interaction vertices [120, 121].

13.2.4 Kinematic Fitting

Kinematic fitting imposes physical constraints on the particles participating in an interaction and thereby improves the measured track momenta and positions. At the same time hypotheses about the interaction and the participating particles can be tested.

13.2.4.1 Lagrange Multiplier Method

The most commonly used method of imposing constraints on the measured tracks is by way of Lagrange multipliers [141]. Let $\tilde{\boldsymbol{q}} = (\tilde{\boldsymbol{q}}_1, \ldots, \tilde{\boldsymbol{q}}_n)^{\mathrm{T}}$ be the unconstrained estimated parameters of a set of n tracks, along with their joint information matrix $\boldsymbol{G} = \mathrm{diag}(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_n) = \boldsymbol{V}^{-1}$. The r functions describing the constraints can be written as $\boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{0}$. Taylor expansion around a suitable point $\breve{\boldsymbol{q}}$ yields the linearized equation

$$\breve{\boldsymbol{g}} + \boldsymbol{D}(\boldsymbol{q} - \breve{\boldsymbol{q}}) = \boldsymbol{0},$$

where $\boldsymbol{D}$ is the Jacobian of $\boldsymbol{g}$ with respect to $\boldsymbol{q}$, evaluated at $\breve{\boldsymbol{q}}$, and $\breve{\boldsymbol{g}} = \boldsymbol{g}(\breve{\boldsymbol{q}})$. The obvious expansion point is $\breve{\boldsymbol{q}} = \tilde{\boldsymbol{q}}$. The constrained track parameters $\bar{\boldsymbol{q}}_i$ are obtained by minimizing the objective function

$$M(\boldsymbol{q}, \boldsymbol{\lambda}) = (\boldsymbol{q} - \tilde{\boldsymbol{q}})^{\mathrm{T}} \boldsymbol{G} (\boldsymbol{q} - \tilde{\boldsymbol{q}}) + 2\boldsymbol{\lambda}^{\mathrm{T}} \left[\breve{\boldsymbol{g}} + \boldsymbol{D}(\boldsymbol{q} - \breve{\boldsymbol{q}}) \right]$$

with respect to $\boldsymbol{q}$ and $\boldsymbol{\lambda}$. $\boldsymbol{\lambda}$ is a vector of r unknowns, the Lagrange multipliers. The solution is

$$\bar{\boldsymbol{q}} = \tilde{\boldsymbol{q}} - \boldsymbol{V}\boldsymbol{D}^{\mathrm{T}}\bar{\boldsymbol{\lambda}}, \quad \text{with} \quad \bar{\boldsymbol{\lambda}} = \boldsymbol{G}_D \left[\breve{\boldsymbol{g}} + \boldsymbol{D}(\tilde{\boldsymbol{q}} - \breve{\boldsymbol{q}}) \right] \quad \text{and} \quad \boldsymbol{G}_D = (\boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^{\mathrm{T}})^{-1}.$$

The covariance matrix $\bar{\boldsymbol{V}}$ and the χ^2 statistic are given by

$$\bar{\boldsymbol{V}} = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{G}_D\boldsymbol{D}\boldsymbol{V}, \quad \chi^2 = \bar{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{G}_D^{-1}\bar{\boldsymbol{\lambda}} = \bar{\boldsymbol{\lambda}}^{\mathrm{T}} \left[\breve{\boldsymbol{g}} + \boldsymbol{D}(\tilde{\boldsymbol{q}} - \breve{\boldsymbol{q}}) \right].$$

If required, the constraint function $\boldsymbol{g}$ can be re-expanded at the new point $\breve{\boldsymbol{q}} = \bar{\boldsymbol{q}}$, and the constrained track parameters can be recomputed.

The Jacobian $\boldsymbol{D}$ depends both on the parametrization of the tracks and on the type of constraint to be imposed. For kinematic constraints it is often convenient to choose a parametrization that uses physically meaningful quantities. In [142] it is proposed to use the four-momentum and a point in space, i.e., $\boldsymbol{q} = (p_x, p_y, p_z, E, x, y, z)$. With this parametrization the following constraints can be formulated in a straightforward manner (for further examples see [142]).

1. Invariant mass constraint. The equation that constrains a track to have an invariant mass m_c is

$$E^2 - {p_x}^2 - {p_y}^2 - {p_z}^2 - {m_c}^2 = 0.$$

Expanding at $\breve{\boldsymbol{q}} = (\breve{p}_x, \breve{p}_y, \breve{p}_z, \breve{E}, \breve{x}, \breve{y}, \breve{z})$ yields

$$\boldsymbol{D} = \left(-2\breve{p}_x \ -2\breve{p}_y \ -2\breve{p}_z \ 2\breve{E} \ 0 \ 0 \ 0\right), \quad \breve{\boldsymbol{g}} = \breve{E}^2 - {\breve{p}_x}^2 - {\breve{p}_y}^2 - {\breve{p}_z}^2 - {m_c}^2.$$

2. Total energy constraint. The equation that constrains a track to have a total energy E_c is

$$E - E_c = 0.$$

It follows that

$$\boldsymbol{D} = \left(0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0\right), \quad \breve{\boldsymbol{g}} = \breve{E} - E_c.$$

3. Total momentum constraint. The equation that constrains a track to have a total momentum p_c is

$$\sqrt{{p_x}^2 + {p_y}^2 + {p_z}^2} - p_c = 0.$$

Expanding at $\breve{\boldsymbol{q}} = (\breve{p}_x, \breve{p}_y, \breve{p}_z, \breve{E}, \breve{x}, \breve{y}, \breve{z})$ yields

$$\boldsymbol{D} = \left(\frac{\breve{p}_x}{\breve{p}} \ \frac{\breve{p}_y}{\breve{p}} \ \frac{\breve{p}_z}{\breve{p}} \ 0 \ 0 \ 0 \ 0\right), \quad \breve{\boldsymbol{g}} = \sqrt{{\breve{p}_x}^2 + {\breve{p}_y}^2 + {\breve{p}_z}^2} - p_c.$$

13.2.4.2 Vertex Constraint

If a vertex constraint is added to the kinematic constraints, the constraint functions depend on the unknown vertex position $\boldsymbol{v}$ and are extended to $\boldsymbol{g}(\boldsymbol{q}, \boldsymbol{v}) = \boldsymbol{0}$. Taylor expansion around a suitable point $(\breve{\boldsymbol{q}}, \breve{\boldsymbol{v}})$ yields the linearized equation

$$\breve{\boldsymbol{g}} + \boldsymbol{D}(\boldsymbol{q} - \breve{\boldsymbol{q}}) + \boldsymbol{E}(\boldsymbol{v} - \breve{\boldsymbol{v}}) = \boldsymbol{0},$$

where $\boldsymbol{E}$ is the Jacobian of $\boldsymbol{g}$ with respect to $\boldsymbol{v}$, evaluated at $\breve{\boldsymbol{v}}$, and $\breve{\boldsymbol{g}} = \boldsymbol{g}(\breve{\boldsymbol{q}}, \breve{\boldsymbol{v}})$. It is assumed that there is prior information about the vertex position, represented by the position $\tilde{\boldsymbol{v}}$ and the covariance matrix $\boldsymbol{C}$. The position $\tilde{\boldsymbol{v}}$ can be used as the expansion point $\breve{\boldsymbol{v}}$.

The constrained track parameters $\bar{\boldsymbol{q}}_i$ and the estimated vertex position $\bar{\boldsymbol{v}}$ are obtained by minimizing the objective function

$$M(\boldsymbol{q}, \boldsymbol{v}, \boldsymbol{\lambda}) = (\boldsymbol{q}-\tilde{\boldsymbol{q}})^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{q}-\tilde{\boldsymbol{q}})+(\boldsymbol{v}-\tilde{\boldsymbol{v}})^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{v}-\tilde{\boldsymbol{v}})+2\boldsymbol{\lambda}^{\mathrm{T}}\left[\breve{\boldsymbol{g}} + \boldsymbol{D}(\boldsymbol{q} - \breve{\boldsymbol{q}}) + \boldsymbol{E}(\boldsymbol{v} - \breve{\boldsymbol{v}})\right]$$

with respect to $\boldsymbol{q}$, $\boldsymbol{v}$, and $\boldsymbol{\lambda}$. The solution is

$$\bar{\boldsymbol{\lambda}} = \boldsymbol{W}\left[\breve{\boldsymbol{g}} + \boldsymbol{D}(\tilde{\boldsymbol{q}} - \breve{\boldsymbol{q}}) + \boldsymbol{E}(\tilde{\boldsymbol{v}} - \breve{\boldsymbol{v}})\right], \quad \bar{\boldsymbol{v}} = \tilde{\boldsymbol{v}} - \boldsymbol{C}\boldsymbol{E}^{\mathrm{T}}\bar{\boldsymbol{\lambda}}, \quad \bar{\boldsymbol{q}} = \tilde{\boldsymbol{q}} - \boldsymbol{V}\boldsymbol{D}^{\mathrm{T}}\bar{\boldsymbol{\lambda}},$$

with $\boldsymbol{W} = \left(\boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{C}\boldsymbol{E}^{\mathrm{T}}\right)^{-1}$. The covariance matrices are

$$\mathrm{Var}(\bar{\boldsymbol{v}}) = \boldsymbol{C}-\boldsymbol{C}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{E}\boldsymbol{C}, \quad \mathrm{Var}(\bar{\boldsymbol{q}}) = \boldsymbol{V}-\boldsymbol{V}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{D}\boldsymbol{V}, \quad \mathrm{Cov}(\bar{\boldsymbol{q}}, \bar{\boldsymbol{v}}) = -\boldsymbol{V}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{E}\boldsymbol{C}.$$

The χ^2 statistic is

$$\chi^2 = \bar{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{W}^{-1}\bar{\boldsymbol{\lambda}} = \bar{\boldsymbol{\lambda}}^{\mathrm{T}}\left[\breve{\boldsymbol{g}} + \boldsymbol{D}(\tilde{\boldsymbol{q}} - \breve{\boldsymbol{q}}) + \boldsymbol{E}(\tilde{\boldsymbol{v}} - \breve{\boldsymbol{v}})\right],$$

with r degrees of freedom, where r is the number of constraint functions. If the vertex constraint is the only constraint imposed on the tracks, the χ^2 has $2n$ degrees of freedom. If there is no prior information about the vertex, the prior vertex position is assigned an infinitely large covariance matrix, and $\boldsymbol{W}$ is replaced by its limiting value:

$$\boldsymbol{W} = \lim_{\boldsymbol{C}\to\infty}\left(\boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{C}\boldsymbol{E}^{\mathrm{T}}\right)^{-1} = \boldsymbol{G}_D - \boldsymbol{G}_D\boldsymbol{E}(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{G}_D\boldsymbol{E})^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{G}_D.$$

The number of degrees of freedom is reduced to $2n - 3$.

13.3 Track Reconstruction in the LHC Experiments

13.3.1 ALICE

ALICE [143] is the experiment at the LHC that is devoted to the physics of high energy ion collisions. Its main goal is to investigate the physics of strongly interacting matter and the quark-gluon plasma at extreme values of energy density and temperature in nucleus-nucleus collisions. Among the four experiments at the LHC, ALICE is equipped with the largest number of subdetectors in order to face the reconstruction complexity of ion physics events. In particular, three subdetectors focus on measuring the passage of charged particles using the bending power of the magnetic field. They are assembled in a cylindrical fashion: the Inner Tracking System (ITS) with six planes of high-resolution silicon pixel, drift, and strip detectors, the cylindrical Time-Projection Chamber (TPC) and the Transition Radiation Detector (TRD). The principal functions of the ITS are the

identification and reconstruction of secondary vertices, the track reconstruction of low-p_T particles and the improvement of the impact parameter and momentum resolution. The TPC is the most important tracking sub-detector. Thanks to its time information, it can provide an efficient and robust tracking also in a very high multiplicity environments (in the order of 10,000 charged particles). Finally, the TRD is also used for tracking in the central region and for improving the p_T resolution at high momentum.

The first step in the track reconstruction in ALICE is the clusterization, which is performed separately for each of the three subdetectors [144]. Tracking then proceeds by determining the preliminary interaction vertex using tracklets defined as lines built with pairs of clusters in the first two layers of the ITS. The preliminary interaction vertex is thus found as a space point to which a maximum number of tracklets converge. In the next step, track finding and fitting is performed in three stages using a inward-outward-inward strategy:

- Initially, tracks in the TPC are searched for using the Kalman filter technique and the outermost layers of the TPC for the seed. A preliminary particle identification is also possible at this stage based on the specific energy loss in the TPC gas. Then, the reconstructed TPC tracks are propagated to the outermost ITS layer and become the seeds for finding tracks in the ITS. In Fig. 13.4 the ITS–TPC matching efficiency as a function of the transverse momentum for 2010–2013 data and Monte Carlo for pp and heavy ion collisions is shown. Finally, the last step is performed in order to recover tracks of particle with p_T down to 80 MeV. It performs a standalone ITS reconstruction with those clusters that were not used in the ITS–TPC tracks.
- All reconstructed tracks are then extrapolated to their point of closest approach to the preliminary interaction vertex, and are extrapolated from the innermost layer to the outermost one. Tracks are refitted by the Kalman filter using the clusters found at the previous stage. After the reconstruction in the TRD subdetectors, the track is matched with a possible TRD tracklet in each of the six TRD layers. In a

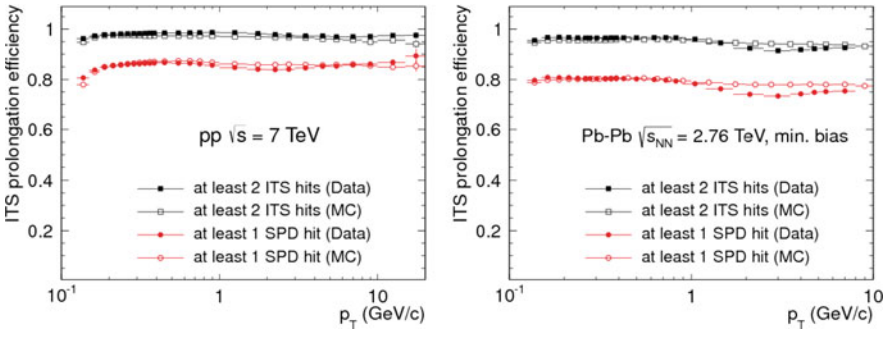


Fig. 13.4 ITS–TPC matching efficiency vs. p_T for data and Monte Carlo for pp (left) for Pb-Pb (right) collisions in the ALICE experiment [144]

similar way, the tracks reaching the time-of-flight (TOF) detector are matched to TOF clusters.
- At the final stage of the track reconstruction, all tracks in both ITS and TPC subdetectors are propagated inwards and refitted one last time to determine the final estimate of the track position, direction, inverse curvature, and its associated covariance matrix.

The final interaction vertex is then re-determined using the all tracks reconstructed in TPC and ITS. The precise vertex fit is performed using track weighting to suppress the contribution of any remaining outliers. For data-taking conditions where a high pileup rate is expected, a more robust version of vertex finding inspired by the algorithm described in [132] is used. It is based on iterative vertex finding and fitting using Tukey bisquare weights to suppress outliers. The algorithm stops when no more vertices are identified in the scan along the beam direction. Once the tracks and the interaction vertex have been found, a search for photon conversions and decays of strange hadrons such as K^0_S and Λ_0 concludes the central-barrel tracking procedure.

13.3.2 ATLAS

ATLAS [145] is the largest of the four LHC experiments, measuring 25 m in diameter and 44 m in length. Its magnet system is composed of a Central Solenoid Magnet with a 2 T field, a Barrel Toroid and an Endcap Toroids with 4 T each. The Inner Detector (ID) is very compact and highly sensitive in order to measure accurately the decay products of each collision. It consists of three different systems of sensors immersed in the solenoid magnetic field: the Pixel Detector, the Semiconductor Tracker (SCT), and the Transition Radiation Tracker (TRT). The Pixel Detector is situated closest to the interaction point and has the highest granularity with about 80 million readout channels. The intrinsic spatial resolution of the Pixel Detector sensors is 10 μm in r–ϕ and 115 μm in z. The SCT is a silicon microstrip detector surrounding the Pixel Detector. It provides eight measurements per track with an overall resolution of 16 μm in r–ϕ and 580 μm in z. In the outermost region, the TRT is placed. It is a light-weight detector composed of proportional gas counters (70% Xe, 27% CO_2 and 3% O_2 straws) embedded in a radiator material and its operational drift radius accuracy is about 130 μm. The TRT contributes both to the track pattern recognition stage, featuring typically around 30 hits per track, and to particle identification.

The basic concepts of the ATLAS track reconstruction are described in [146, 147]. The tracking in the ID consists of two principal sequences: an initial inside-out tracking, and a subsequent outside-in tracking. Inside-out tracking starts with space point formation in the silicon part of the ID. Using the space points, track seeds are generated with or without a constraint on the longitudinal vertex position. The seeds are then followed through the SCT by a combinatorial Kalman filter/smoother.

After ambiguity solving, the remaining track candidates are extended into the TRT. Outside-in tracking first finds track segments in the TRT, using a Hough transform of the straw centers. A Kalman filter/smoother using also the drift times builds the final track segments. These track segments are then extrapolated back into the SCT and the Pixel Detector. Muons are reconstructed in the ID like any other charged particles; for the standalone reconstruction of muons in the muon system and the combined reconstruction, see [148].

Based on the experience gained in Run 1, several improvements to track reconstruction were made for Run 2 [149]. For example, the tracking was adapted to the new insertable B-layer (IBL) [150], and track reconstruction in dense environments (TIDE) was optimized [151]. This included an artificial neural network based approach to identify pixel clusters created by multiple charged particles. The effect of these two developments is shown in Fig. 13.5. In Fig. 13.5a the transverse impact parameter as a function of track momentum resolution is shown for data taken in 2015 at 13 TeV with the inclusion of the IBL information and for data in 2012 at 8 TeV without the IBL. The data in 2015 was collected with a minimum bias trigger. The data in 2012 is derived from a mixture of jet, tau and missing E_{T} triggers [150]. Figure 13.5b shows the improvement of the track reconstruction efficiency in the jet core due to the TIDE optimization [151].

ATLAS track reconstruction efficiency as a function of pseudorapidity and transverse momentum with simulated data at a center-of-mass energy of 13 TeV is shown in Fig. 13.6 [152].

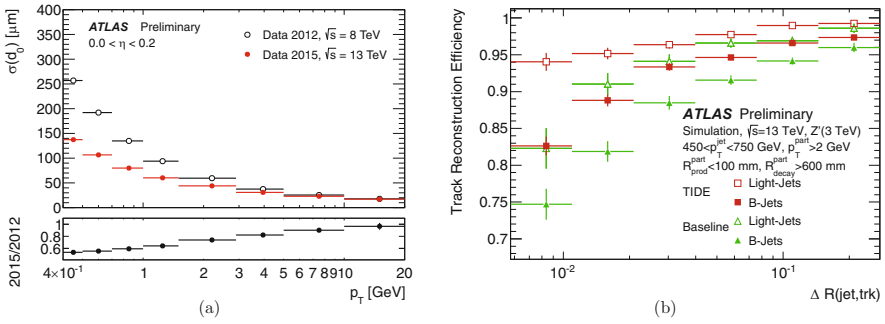


Fig. 13.5 (**a**) Upper panel: unfolded transverse impact parameter resolution measured from data in 2015 at 13 TeV with the Inner Detector including the IBL, as a function of track p_{T} for values of $0.0 < \eta < 0.2$, compared to that measured from data in 2012 at 8 TeV [150]; lower panel: ratio of the resolution in 2015 over the resolution in 2012. (**b**) Improvement of the track reconstruction efficiency due to the TIDE optimization, as a function of the angular distance of the particle from the jet axis. The track selection is explained in [151]

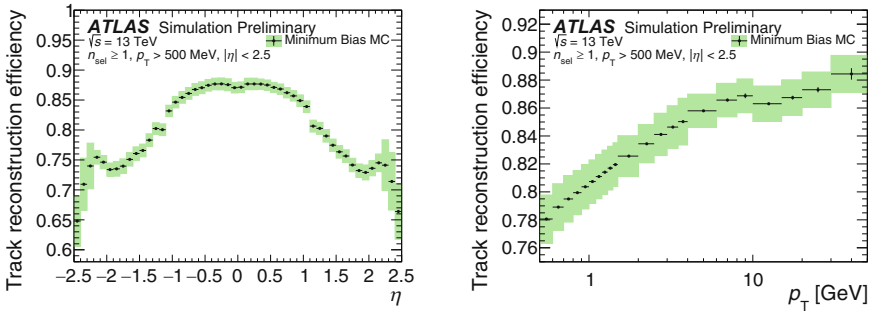


Fig. 13.6 The track reconstruction efficiency (**a**) as a function of pseudorapidity and (**b**) as a function of transverse momentum, as predicted by Pythia 8 A2 simulation. The statistical uncertainties are shown as black lines, the total uncertainties as green shaded areas [152]

13.3.3 CMS

CMS [153], together with ATLAS, is one of the two general-purpose experiments at the LHC. Its main distinguishing feature is a 3.8 T superconducting solenoid. With a length of 13 m and a diameter of 6 m, it provides a high bending power to precisely measure the momentum of charged particles. The solenoid magnetic field lines run parallel to the beam direction in the central region, where the tracking system is placed. The tracking system is designed to provide a precise and efficient measurement of particle trajectories using position-sensitive detectors. The CMS tracker is a silicon-based system [154]. It splits into two parts, the Pixel Tracker and the Strip Tracker and covers a pseudorapidity range up to $|\eta| = 2.5$. The Pixel Tracker is the innermost CMS detector sub-system and is composed of 66 million silicon pixels with dimensions $100 \times 250 \times 250\,\mu\text{m}$, covering a total area of about $1\,\text{m}^2$. In the barrel layers the magnetic field induces a Lorentz angle which increases charge sharing between neighbouring pixels. Charge sharing in conjunction with analog readout allows to achieve $10\,\mu\text{m}$ position resolution for the (r, ϕ) coordinate and $15\,\mu\text{m}$ in the z direction. The pixel detectors in the forward direction are tilted at an angle of 20° to induce charge sharing which allows to achieve $15\,\mu\text{m}$ and $20\,\mu\text{m}$ resolution respectively. This resolution is not only necessary for a precise track reconstruction, but also for the determination of both the vertices produced in the primary interaction and the decay vertices of short-lived particles.

The Strip Tracker constitutes the outer part of the tracking system. Its basic building blocks are silicon strip modules. Each module is equipped with one or two silicon sensors and a so-called Front-End hybrid containing readout electronics. In total, the CMS silicon strip tracker has 9.3 million strips and covers $198\,\text{m}^2$ of active silicon area. The resolution in (r, ϕ) is $\simeq 30\,\mu\text{m}$ in all layers. The inner layers of the strip tracker are equipped with double-sided sensors, one side of which is rotated by a stereo angle of 100 mrad, achieving a resolution along the z coordinate

of about 230 μm and allowing the reconstruction of the hit position in 3-D. In the outer layers the sensors are single-sided, and the z resolution can be approximated by the strip length over $\sqrt{12}$, or about 15 mm. In order to maintain excellent tracking performance until the Long Shutdown 3 of LHC, the Pixel Tracker was replaced in the year-end technical stop of 2016/2017 with a new Pixel Tracker composed of four barrel layers and six forward disks providing four-hit pixel coverage up to $|\eta| = 2.5$. After the Long Shutdown 3, the High Luminosity phase of the LHC (HL-LHC) is scheduled where the accelerator will provide an unprecedented instantaneous luminosity of $5 - 7.5 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. In [155] the new CMS silicon tracker and its tracking and vertexing performance for different event types, pileup scenarios and detector geometries are presented.

The CMS track reconstruction algorithm is based on an iterative approach [156]. The main idea is to search for easier-to-find tracks first, to mask the hits associated to the found tracks, and to proceed to the next iteration. In this way the combinatorial problem is reduced, and the search for more difficult classes of tracks is simplified. Moreover, this approach introduces the possibility of developing special iterations that can improve track reconstruction in high-density environments such as jets, or to use the information from other subsystems such as muon chambers and calorimeters. In each iteration, the Combinatorial Track Finder is run. It can be divided into four different steps:

1. **Seed generation**: Using the information of three or four hits, the trajectory parameters and the corresponding uncertainties of the initial track candidates are computed.
2. **Track finding**: Staring from the seed, the current trajectory parameters and their uncertainties are extrapolated to the next layer and compatible hits are found. Each of them is added to a clone of the track candidate. Each of these candidates is again extrapolated to the next layer and compatible hits are found. This procedure is repeated for each candidate until there is more than one missing hit or the extrapolation does not find another tracker layer.
3. **Track fitting**: A Kalman filter or a Gaussian-sum filter/smoother is performed to obtain the final estimate of the track parameters at the interaction point exploiting the full trajectory information.
4. **Track selection**: Tracks are grouped in classes according to different track quality criteria.

As an example, the twelve tracking iterations foreseen for 2017 data taking is listed in Table 13.3 [157]. The main difference between iterations is the configuration of the seed generation and the target tracks.

Figure 13.7 shows the tracking efficiency, using a standard sample of $\mathrm{t\bar{t}}$ events simulated with $\sqrt{s} = 13\,\mathrm{TeV}$ with different superimposed pileup conditions. The contribution of different iterations for 2017 track reconstruction is also shown as a function of the p_T of the simulated particle. It can be seen how iterations targeting low-p_T tracks are more efficient in the region between 100 and 500 MeV.

Table 13.3 List of different tracking iterations used after the Pixel Tracker upgrade with the corresponding seeding configuration used and target tracks [157]

Iteration	Step name	Seeding	Target tracks
0	Initial	Pixel quadruplets	Prompt, high p_T
1	LowPtQuad	Pixel quadruplets	Prompt, low p_T
2	HighPtTriplet	Pixel triplets	Prompt, high p_T recovery
3	LowPtTriplet	Pixel triplets	Prompt, low p_T
4	DetachedQuad	Pixel quadruplets	From b hadron decay, $r \leq 5$ cm
5	DetachedTriplet	Pixel triplets	From b hadron decay, $r \leq 10$ cm
6	MixedTriplet	Pixel+strip triplets	Displaced, $r \leq 7$ cm
7	PixelLess	Inner strip pairs	Displaced, $r \leq 25$ cm
8	TobTec	Outer strip pairs	Displaced, $r \leq 60$ cm
9	JetCore	Pixel pairs in jets	High-p_T jets
10	Muon inside-out	Muon-tagged tracks	Muons
11	Muon outside-in	Standalone muons	Muons

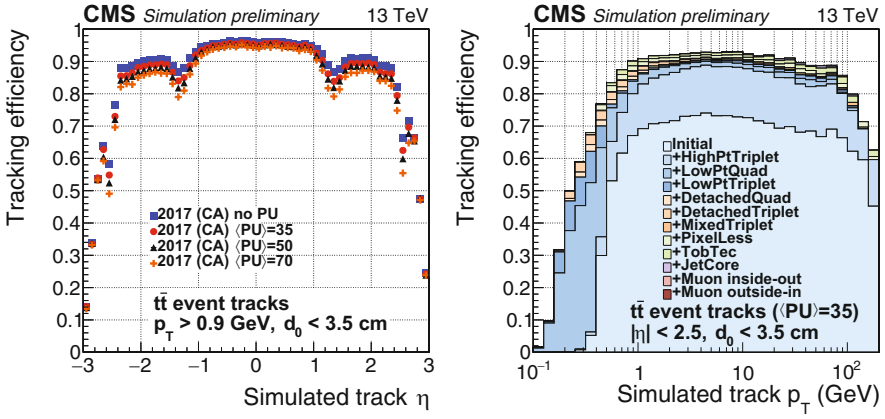


Fig. 13.7 Track reconstruction efficiency as a function of simulated track pseudorapidity for 2017 tracker at different pileup conditions (left) and cumulative contributions to the overall tracking performance from the twelve iterations in 2017 track reconstruction shown as a function of the simulated track p_T (right) [157]. The 2017 tracking reconstruction includes the Cellular Automaton-based Hit Chain-Maker (CA) seeding [158]

Figure 13.8 shows the muon tracking efficiency and the corresponding ratios between real and simulated data for 2016 collisions data coming from the Z resonance using the tag and probe method. The measured track efficiency as a function of $|\eta|$ is found to be between 99.5% and 100% for the collection including all tracks. It degrades, however, with increasing number of primary vertices.

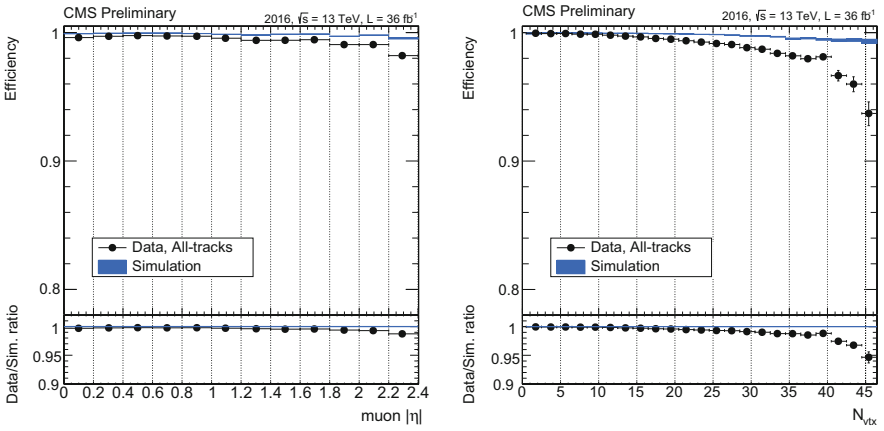


Fig. 13.8 Data (black dots) and simulation (rectangles) tracking efficiency and respective ratio for muons coming from the Z decay as a function of the absolute pseudorapidity of the probe muon (left) and the number of primary vertices (right). The data are based on an integrated luminosity of 36 fb^{-1} [155]

13.3.4 LHCb

As its name indicates, LHCb [159] focuses on physics involving bottom quarks and investigates CP violation phenomena. These studies require the measurement of the rare decays of B_d, B_s, and D mesons which are produced with a large cross-section at the LHC. Given the fact that b hadrons are predominantly produced in the forward or backward cone, the LHCb experiment is a single-arm spectrometer in contrast to the other three experiments. In order to exploit this large number of b hadrons, it requires a robust and flexible trigger and a data acquisition that allows high bandwidth data taking and provides powerful online data processing. Furthermore, superior vertex and momentum resolution are crucial to study the rapidly oscillating $B_s - \overline{B_s}$ meson system. LHCb is thus equipped with the highly sophisticated silicon microstrip detector close to the interaction point, the Vertex Locator (VELO). It can be moved to a distance of only 7 mm from the proton beams and measures the position of the primary vertices and the impact parameters of the track with extremely high precision. A further silicon microstrip detector, the Tracker Turicensis (TT) is placed before the dipole magnet. Its task is to improve the momentum resolution of reconstructed tracks and reject pairs of tracks that in reality belong to the same particle. The magnet is placed behind the TT. It bends the flight path of the particles in the $x - z$ plane and therefore allows the determination of their momenta. The tracking system is completed by the T stations (T1-T2-T3), which, together with the information from the VELO, determine the momentum and flight direction of the particles. The T stations are composed of silicon microstrip sensors close to the beam pipe and by straw tubes in the outer regions.

Track reconstruction uses hits in the VELO, TT and T stations. Depending on which detectors are crossed, different track types are defined [160, 161]:

- **Long tracks** traverse the full tracking system. They have hits in both the VELO and the T stations, and optionally in TT. They are the most important set of tracks for physics analyses.
- **Upstream tracks** pass only through the VELO and TT stations. In general their momentum is too low to traverse the magnet and reach the T stations.
- **Downstream tracks** pass only through the TT and T stations. They are important for the reconstruction of long-lived particles that decay outside the VELO acceptance.
- **VELO tracks** pass only through the VELO. These tracks are particularly important in the primary vertex reconstruction.
- **T tracks** pass only through the T stations. Like the downstream tracks, they are useful for particle identification in the Ring Imaging Cherenkov detectors.

Reconstruction of long tracks starts in the VELO. There are two complementary algorithms to add information from the downstream tracking stations to these VELO tracks. The first one combines the VELO tracks with information from the T stations. The second one combines the VELO tracks with track segments found after the magnet in the T stations, using a standalone track finding algorithm. The candidate tracks found by each algorithm are then combined, removing duplicates, to form the final set of long tracks used for analysis. Finally, hits in the TT consistent with the extrapolated trajectories of each track are added to improve their momentum determination.

Downstream tracks are found starting with T tracks, extrapolating them through the magnetic field and searching for corresponding hits in the TT. Upstream tracks are found by extrapolating VELO tracks to the TT where matching hits are then added in a procedure similar to that used by the downstream tracking. At least three TT hits are required to be present by these algorithms.

The found tracks are fitted using a Kalman filter, taking into account multiple scattering and energy loss due to ionisation. The χ^2-statistic of the fit is used to determine the quality of the reconstructed track. If two or more tracks have many hits in common, only the one with most hits is kept.

The track reconstruction efficiency for the 2012 and the 2015 data as a function of the momentum can be seen in Fig. 13.9 [162]. The results of the two periods are compatible.

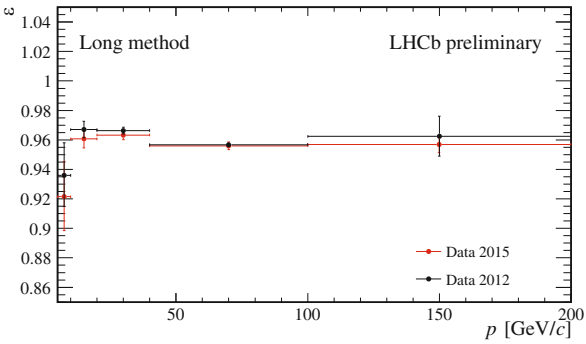


Fig. 13.9 LHCb track reconstruction efficiency for the 2012 and the 2015 data as a function of the momentum. The efficiency is computed using the "Long method", described in [161]

13.4 Conclusion

An overview of current methods in track and vertex reconstruction and alignment has been presented. Many of them have been developed in response to the requirements of the current experimental program at the Large Hadron Collider. The most difficult challenges are:

- Reliable reconstruction of signal events over a large background of non-signal events, pileup events, and low-momentum tracks;
- Reliable reconstruction of secondary vertices with very short distances from the primary vertex;
- Precise alignment of a large number of sensors.

Every experiment has to meet these challenges on its own terms. The outlines of the solutions found by the four major LHC experiments are described in Sect. 13.3 and, in more detail, in the references given there. In addition, the repertory of the methods discussed in this contribution can certainly not lay claim to completeness. We have tried to select widely applicable methods, thereby neglecting by necessity many experiment specific adaptations, improvements and innovations, for which we again refer to the references.

References

1. R. Mankel, "Pattern recognition and event reconstruction in particle physics experiments," *Rept. Prog. Phys.*, vol. 67, p. 553, 2004.
2. M. Hansroul, H. Jeremie, and D. Savard, "Fast circle fit with the conformal mapping method," *Nucl. Instrum. Meth.*, vol. A270, no. 2, p. 498, 1988.
3. P. Hough, "Machine analysis of bubble chamber pictures," in *Proceedings of the International Conference on High Energy Accelerators and Instrumentation*, (CERN, Geneva), p. 554, 1959.
4. H. Kälviäinen, P. Hirvonen, L. Xu, and E. Oja, "Probabilistic and non-probabilistic Hough transforms: overview and comparisons," *Image and Vision Computing*, vol. 13, no. 4, p. 239, 1995.

5. T. Alexopoulos, M. Bachtis, E. Gazis, and G. Tsipolitis, "Implementation of the Legendre Transform for track segment reconstruction in drift tube chambers," *Nucl. Instrum. Meth.*, vol. A592, pp. 456–462, 2008.
6. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci.*, vol. 79, p. 2554, 1982.
7. C. Peterson, "Track Finding With Neural Networks," *Nucl. Instrum. Meth.*, vol. A279, p. 537, 1989.
8. B. Denby, "Neural Networks and Cellular Automata in Experimental High-energy Physics," *Comput. Phys. Commun.*, vol. 49, p. 429, 1988.
9. C. Peterson and J. Anderson, "A Mean Field Theory Learning Algorithm for Neural Networks," *Complex Systems*, vol. 2, p. 995, 1987.
10. M. Diehl, M. Junger, R. Frühwirth, and J. Scherzer, "Global optimization for track finding," *Nucl. Instrum. Meth.*, vol. A389, p. 180, 1997.
11. G. Stimpfl-Abele and L. Garrido, "Fast track finding with neural nets," *Comput. Phys. Commun.*, vol. 64, p. 46, 1991.
12. S. Baginian, A. Glazov, I. Kisel, E. Konotopskaya, V. Neskoromnyi, and G. Ososkov, "Tracking by a modified rotor model of neural network," *Comput. Phys. Commun.*, vol. 79, p. 165, 1994.
13. A. Badalà, R. Barbera, G. Lo Re, A. Palmeri, G. S. Pappalardo, A. Pulvirenti, and F. Riggi, "Neural tracking in ALICE," *Nucl. Instrum. Meth.*, vol. A502, pp. 503–506, 2003.
14. A. Pulvirenti, A. Badala, R. Barbera, G. Lo Re, A. Palmeri, G. S. Pappalardo, and F. Riggi, "Neural tracking in the ALICE Inner Tracking System," *Nucl. Instrum. Meth.*, vol. A533, pp. 543–559, 2004.
15. A. Badalà, R. Barbera, G. Lo Re, A. Palmeri, G. S. Pappalardo, A. Pulvirenti, and F. Riggi, "Combined tracking in the ALICE detector," *Nucl. Instrum. Meth.*, vol. A534, p. 211, 2004.
16. M. Gyulassy and M. Harlander, "Elastic tracking and neural network algorithms for complex pattern recognition," *Comput. Phys. Commun.*, vol. 66, p. 31, 1991.
17. M. Gyulassy and M. Harlander, "High resolution multiparticle tracking without preprocessing via elastic tracking," *Nucl. Instrum. Meth.*, vol. A316, p. 238, 1992.
18. M. Ohlsson, C. Peterson, and A. Yuille, "Track finding with deformable templates: The Elastic arms approach," *Comput. Phys. Commun.*, vol. 71, p. 77, 1992.
19. M. Ohlsson, "Extensions and explorations of the elastic arms algorithm," *Comput. Phys. Commun.*, vol. 77, p. 19, 1993.
20. R. Durbin and D. Willshaw, "An analogue approach to the travelling salesman," *Nature*, vol. 326, no. 16, p. 689, 1987.
21. D. Bui, T. Greenshaw, and G. Schmidt, "A combination of an elastic net and a Hopfield net to solve the segment linking problem in the forward tracker of the H1 detector at HERA," *Nucl. Instrum. Meth.*, vol. A389, p. 184, 1997.
22. I. Kisel and V. Kovalenko, "Elastic net for broken multiple scattered tracks," *Comput. Phys. Commun.*, vol. 98, p. 45, 1996.
23. I. Kisel *et al.*, "Cellular automaton and elastic net for event reconstruction in the NEMO-2 experiment," *Nucl. Instrum. Meth.*, vol. A387, p. 433, 1997.
24. R. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. 1, p. 35, 1960.
25. D. Catlin, *Estimation, Control, and the Discrete Kalman Filter*. NewYork: Springer, 1989.
26. R. Frühwirth, "Application of Kalman filtering to track and vertex fitting," *Nucl. Instrum. Meth.*, vol. A262, p. 444, 1987.
27. P. Billoir, "Progressive track recognition with a Kalman like fitting procedure," *Comput. Phys. Commun.*, vol. 57, p. 390, 1989.
28. R. Mankel, "A Concurrent track evolution algorithm for pattern recognition in the HERA-B main tracking system," *Nucl. Instrum. Meth.*, vol. A395, p. 169, 1997.
29. A. Glazov, I. Kisel, E. Konotopskaya, and G. Ososkov, "Filtering tracks in discrete detectors using a cellular automaton," *Nucl. Instrum. Meth.*, vol. A329, p. 262, 1993.

30. R. Frühwirth and A. Strandlie, "Application of adaptive filters to track finding," *Nucl. Instrum. Meth.*, vol. A559, p. 162, 2006.
31. A. Strandlie and R. Frühwirth, "Reconstruction of charged tracks in the presence of large amounts of background and noise," *Nucl. Instrum. Meth.*, vol. A566, p. 157, 2006.
32. A. Strandlie and R. Frühwirth, "Track and vertex reconstruction: From classical to adaptive methods," *Rev. Mod. Phys.*, vol. 82, pp. 1419–1458, May 2010.
33. H. Wind, "Evaluating a magnetic field component from boundary observations only," *Nucl. Instrum. Meth.*, vol. 84, p. 117, 1970.
34. M. Aleksa *et al.*, "Measurement of the ATLAS solenoid magnetic field," *JINST*, vol. 3, p. P04003, 2008.
35. R. Frühwirth, "Prepating Magnetic Field Measurements," *Comput. Phys. Commun.*, vol. 22, p. 223, 1981.
36. N. Amapane, V. Andreev, V. Drollinger, V. Karimäki, V. Klyukhin, and T. Todorov, "Volume-based representation of the magnetic field," in *Computing in high energy physics and nuclear physics. Proceedings, Conference, CHEP'04, Interlaken, Switzerland, September 27-October 1, 2004*, p. 310, 2004.
37. R. Frühwirth *et al.*, *Data Analysis Techniques for High-Energy Physics*. Cambridge: Cambridge University Press, 2 ed., 2000.
38. E. Lund, L. Bugge, I. Gavrilenko, and A. Strandlie, "Track parameter propagation through the application of a new adaptive Runge-Kutta-Nystroem method in the ATLAS experiment," *JINST*, vol. 4, p. P04001, 2009.
39. A. Strandlie and W. Wittek, "Derivation of Jacobians for the propagation of covariance matrices of track parameters in homogeneous magnetic fields," *Nucl. Instrum. Meth.*, vol. A566, no. 2, p. 687, 2006.
40. J. Myrheim and L. Bugge, "A Fast Runge-Kutta Method for Fitting Tracks in a Magnetic Field," *Nucl. Instrum. Meth.*, vol. 160, p. 43, 1979.
41. L. Bugge and J. Myrheim, "Tracking and Track fitting," *Nucl. Instrum. Meth.*, vol. 179, p. 365, 1981.
42. R. Frühwirth and M. Regler, "On the quantitative modelling of core and tails of multiple scattering by Gaussian mixtures," *Nucl. Instrum. Meth.*, vol. A456, no. 3, p. 369, 2001.
43. J. Jackson, *Classical Electrodynamics*. John Wiley & Sons, 2007.
44. C. Patrignani *et al.*, "Review of Particle Physics," *Chin. Phys.*, vol. C40, no. 10, p. 100001, 2016.
45. H. Eichinger and M. Regler, "Review of track-fitting methods in counter experiments," Tech. Rep. CERN-81-06, CERN, 1981.
46. P. Avery, "Applied fitting theory V: track fitting using the Kalman filter," Tech. Rep. CLEO Note CBX92-39, Cornell University, 1992.
47. V. Innocente and E. Nagy, "Trajectory fit in presence of dense materials," *Nucl. Instrum. Meth.*, vol. A324, p. 297, 1993.
48. H. Bethe and W. Heitler, "On the Stopping of fast particles and on the creation of positive electrons," *Proc. Roy. Soc. Lond.*, vol. A146, p. 83, 1934.
49. R. Frühwirth, "A Gaussian-mixture approximation of the Bethe–Heitler model of electron energy loss by bremsstrahlung," *Comput. Phys. Commun.*, vol. 154, no. 2, p. 131, 2003.
50. W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC," *J. Phys. G: Nuclear and Particle Physics*, vol. 31, no. 9, p. N9, 2005.
51. V. Kartvelishvili, "Electron bremsstrahlung recovery in ATLAS," *Nucl. Phys. Proc. Suppl.*, vol. 172, p. 208, 2007.
52. N. Chernov and G. Ososkov, "Effective Algorithms of Circle Fitting," *Comput. Phys. Commun.*, vol. 33, p. 329, 1984.
53. V. Karimäki, "Effective circle fitting for particle trajectories," *Nucl. Instrum. Meth.*, vol. A305, p. 187, 1991.
54. A. Strandlie, J. Wroldsen, R. Frühwirth, and B. Lillekjendlie, "Particle tracks fitted on the Riemann sphere," *Comput. Phys. Commun.*, vol. 131, p. 95, 2000.

55. A. Strandlie and R. Frühwirth, "Error analysis of the track fit on the Riemann sphere," *Nucl. Instrum. Meth.*, vol. A480, p. 734, 2002.
56. A. Strandlie, J. Wroldsen, and R. Frühwirth, "Treatment of multiple scattering with the generalized Riemann sphere track fit," *Nucl. Instrum. Meth.*, vol. A488, p. 332, 2002.
57. P. Laurikainen, W. Moorhead, and W. Matt, "Least squares fit of bubble chamber tracks taking into account multiple scattering," *Nucl. Instrum. Meth.*, vol. 98, p. 349, 1972.
58. C. Kleinwort, "General broken lines as advanced track fitting method," *Nucl. Instrum. Meth.*, vol. A673, p. 107, 2012.
59. P. Billoir, R. Frühwirth, and M. Regler, "Track Element Merging Strategy and Vertex Fitting in Complex Modular Detectors," *Nucl. Instrum. Meth.*, vol. A241, p. 115, 1985.
60. P. Billoir, "Track Fitting With Multiple Scattering: A New Method," *Nucl. Instrum. Meth.*, vol. A225, pp. 352–366, 1984.
61. R. Harr, "Calculation of track and vertex errors for detector design studies," *IEEE Trans. Nucl. Sci.*, vol. 42, p. 134, 1995.
62. R. Frühwirth, "Track fitting with non-Gaussian noise," *Comput. Phys. Commun.*, vol. 100, p. 1, 1997.
63. A. Strandlie and R. Frühwirth, "Discrimination between different types of material in track reconstruction with a Gaussian-sum filter," *IEEE Trans. Nucl. Sci.*, vol. 53, p. 3842, 2006.
64. R. Frühwirth and A. Strandlie, "Track fitting with ambiguities and noise: A study of elastic tracking and nonlinear filters," *Comput. Phys. Commun.*, vol. 120, p. 197, 1999.
65. A. Strandlie and R. Frühwirth, "Adaptive multitrack fitting," *Comput. Phys. Commun.*, vol. 133, p. 34, 2000.
66. R. Frühwirth, A. Strandlie, M. Winkler, and T. Todorov, "Recent results on adaptive track and multitrack fitting," *Nucl. Instrum. Meth.*, vol. A502, p. 702, 2003.
67. S. Catani, Y. Dokshitzer, M. Olsson, G. Turnock, and B. Webber, "New clustering algorithm for multi-jet cross-sections in $e^+ e^-$ annihilation," *Phys. Lett.*, vol. B269, p. 432, 1991.
68. S. Catani, Y. Dokshitzer, M. Seymour, and B. Webber, "Longitudinally invariant $k_\perp$ clustering algorithms for hadron-hadron collisions," *Nucl. Phys.*, vol. B406, p. 187, 1993.
69. S. Moretti, L. Lönnblad, and T. Sjöstrand, "New and old jet clustering algorithms for electron-positron events," *J. High Energy Phys.*, vol. 08, p. 001, 1998.
70. W. Bartel *et al.*, "Experimental Studies on Multi-Jet Production in $e^+ e^-$ Annihilation at PETRA Energies," *Z. Phys.*, vol. C33, p. 23, 1986. [,53(1986)].
71. Y. Dokshitzer, G. Leder, S. Moretti, and B. Webber, "Better jet clustering algorithms," *J. High Energy Phys.*, vol. 08, p. 001, 1997.
72. M. Seymour and C. Tevlin, "A Comparison of two different jet algorithms for the top mass reconstruction at the LHC," *J. High Energy Phys.*, vol. 11, p. 052, 2006.
73. M. Cacciari and G. Salam, "Dispelling the N^3 myth for the k_t jet-finder," *Phys. Lett.*, vol. B641, p. 57, 2006.
74. T. Sjöstrand, "The Lund Monte Carlo for e+ e- Jet Physics," *Comput. Phys. Commun.*, vol. 28, p. 229, 1983.
75. S. Bethke, Z. Kunszt, D. Soper, and W. Stirling, "New jet cluster algorithms: Next-to-leading order QCD and hadronization corrections," *Nucl. Phys.*, vol. B370, p. 310, 1992. [Erratum: Nucl. Phys.B523,681(1998)].
76. J. Dorfan, "A Cluster Algorithm for the Study of Jets in High-Energy Physics," *Z. Phys.*, vol. C7, p. 349, 1981.
77. J. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Amer. Math. Soc.*, vol. 7, no. 1, p. 48, 1956.
78. L. Angelini, G. Nardulli, L. Nitti, M. Pellicoro, D. Perrino, and S. Stramaglia, "Deterministic annealing as a jet clustering algorithm in hadronic collisions," *Phys. Lett.*, vol. B601, p. 56, 2004.
79. S. Chekanov, "A New jet algorithm based on the k-means clustering for the reconstruction of heavy states from jets," *Eur. Phys. J.*, vol. C47, p. 611, 2006.
80. S.-L. Blyth *et al.*, "A Cone jet-finding algorithm for heavy-ion collisions at LHC energies," *J. Phys.*, vol. G34, p. 271, 2007.

81. F. Tarrade, "Reconstruction and identification of hadronic tau decays in ATLAS," *Nucl. Phys. Proc. Suppl.*, vol. 169, p. 357, 2007.
82. L. Barbone, N. De Filippis, O. L. Buchmüller, F. P. Schilling, T. Speer, and P. Vanlaer, "Impact of CMS silicon tracker misalignment on track and vertex reconstruction," *Nucl. Instrum. Meth.*, vol. A566, p. 45, 2006.
83. S. Blusk, O. Buchmüller, A. Jacholkowski, T. Ruf, J. Schieck, and S. Viret, eds., *Proceedings of the first LHC Detector Alignment Workshop, CERN, Geneva, Switzerland, 4–6 September 2006*, 2007.
84. V. Blobel, "Software alignment for tracking detectors," *Nucl. Instrum. Meth.*, vol. A566, p. 5, 2006.
85. W. Wiedenmann, "Alignment of the ALEPH tracking devices," *Nucl. Instrum. Meth.*, vol. A323, p. 213, 1992.
86. A. Andreazza and E. Piotto, "The Alignment of the DELPHI Tracking Detectors," Tech. Rep. DELPHI 99–153 TRACK 94, CERN, 1999.
87. A. Sopczak, "Alignment of the central D0 Detector," *Nucl. Instrum. Meth.*, vol. A566, p. 142, 2006.
88. Y. Fisyak *et al.*, "Overview of the inner silicon detector alignment procedure and techniques in the RHIC/STAR experiment," *J. Phys. Conf. Ser.*, vol. 119, p. 032017, 2008.
89. D. Brown, A. Gritsan, Z. Guo, and D. Roberts, "Local Alignment of the BABAR Silicon Vertex Tracking Detector," *Nucl. Instrum. Meth.*, vol. A603, p. 467, 2009.
90. P. Schleper, G. Steinbrück, and M. Stoye, "Alignment of the CMS silicon tracker using Millepede II," *J. Phys. Conf. Ser.*, vol. 119, p. 032040, 2008.
91. M. Gersabeck, "Alignment of the LHCb Vertex Locator," *Nucl. Instrum. Meth.*, vol. A598, p. 71, 2009.
92. S. Gonzalez-Sevilla, "Track-based alignment of the ATLAS inner detector," *J. Phys. Conf. Ser.*, vol. 119, p. 032019, 2008.
93. E. Widl and R. Frühwirth, "A large-scale application of the Kalman alignment algorithm to the CMS tracker," *J. Phys. Conf. Ser.*, vol. 119, p. 032038, 2008.
94. S. Chatrchyan *et al.*, "Alignment of the CMS tracker with LHC and cosmic ray data," *JINST*, vol. 9, p. P06009, 2014.
95. G. Mittag, "Alignment of the CMS Tracker: Latest results from LHC Run-II," *J. Phys. Conf. Ser.*, vol. 898, no. 4, p. 042014, 2017.
96. J. Schieck, "Track-based alignment for the ATLAS Inner Detector Tracking System," *JINST*, vol. 7, p. C01012, 2012.
97. G. Ripellino, "The alignment of the ATLAS Inner Detector in Run-2," *PoS*, vol. LHCP2016, p. 196, 2016.
98. M. Martinelli, "Novel real-time alignment and calibration of the LHCb detector in Run2," *J. Phys. Conf. Ser.*, vol. 898, no. 3, p. 032039, 2017.
99. J. Amoraal, "Alignment of the LHCb detector with Kalman filter fitted tracks," *J. Phys. Conf. Ser.*, vol. 219, p. 032028, 2010.
100. A. Dainese, "Alignment of the ALICE tracking detectors," *PoS*, vol. VERTEX2009, p. 021, 2009.
101. V. Karimäki, T. Lampen, and F. Schilling, "The HIP algorithm for track based alignment and its application to the CMS pixel detector," Tech. Rep. CMS-NOTE-2006-018, CERN, 2006.
102. V. Blobel and C. Kleinwort, "A New method for the high precision alignment of track detectors," in *Advanced Statistical Techniques in Particle Physics. Proceedings, Conference, Durham, UK, March 18–22, 2002*, 2002.
103. Y. Saad and M. Schultz, "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal on scientific and statistical computing*, vol. 7, no. 3, p. 856, 1986.
104. E. Widl, R. Frühwirth, and W. Adam, "A Kalman filter for track-based alignment," tech. rep., CERN, 2006.
105. R. Frühwirth, T. Todorov, and M. Winkler, "Estimation of detector alignment parameters using the Kalman filter with annealing," *J. Phys.*, vol. G29, p. 561, 2003.

106. R. Gluckstern, "Uncertainties in track momentum and direction, due to multiple scattering and measurement errors," *Nucl. Instrum. Meth.*, vol. 24, p. 381, 1963.
107. M. Regler and R. Frühwirth, "Generalization of the Gluckstern formulas. I: Higher orders, alternatives and exact results," *Nucl. Instrum. Meth.*, vol. A589, p. 109, 2008.
108. R. Frühwirth and A. Beringer, "JDOT — a Java detector optimization tool," in *Nuclear Science Symposium Conference Record, 2008. NSS'08. IEEE*, pp. 3483–3487, IEEE, 2008.
109. W. Innes, "Some formulas for estimating tracking errors," *Nucl. Instrum. Meth.*, vol. A329, p. 238, 1992.
110. W. Innes, "TRACKERR: A Program for calculating tracking errors," Tech. Rep. SLAC-BABAR-NOTE-121, SLAC, 1993.
111. M. Regler, M. Valentan, and R. Frühwirth, "The LiC detector toy program," *Nucl. Instrum. Meth.*, vol. A581, p. 553, 2007.
112. S. Masciocchi, "Experience with HERA-B vertexing," *Nucl. Instrum. Meth.*, vol. A462, p. 220, 2001.
113. E. Chabanat, J. D'Hondt, N. Estre, R. Frühwirth, K. Prokofiev, T. Speer, P. Vanlaer, and W. Waltenberger, "Vertex reconstruction in CMS," *Nucl. Instrum. Meth.*, vol. A549, p. 188, 2005.
114. W. Erdmann, "Vertexing in the H1 experiment," *Nucl. Instrum. Meth.*, vol. A560, p. 89, 2006.
115. M. Costa, "Vertex and track reconstruction in ATLAS," *Nucl. Instrum. Meth.*, vol. A582, p. 785, 2007.
116. W. Adam, "Track and vertex reconstruction in CMS," *Nucl. Instrum. Meth.*, vol. A582, p. 781, 2007.
117. W. Waltenberger, *Development of vertex finding and vertex fitting algorithms for CMS*. PhD thesis, TU Wien, 2004.
118. D. Jackson, "A Topological vertex reconstruction algorithm for hadronic jets," *Nucl. Instrum. Meth.*, vol. A388, p. 247, 1997.
119. S. Hillert, "ZVMST: A Minimum spanning tree-based vertex finder," tech. rep., 2008.
120. W. Waltenberger and F. Moser, "RAVE — an Open, Extensible, Detector-Independent Toolkit for Reconstruction of Interaction Vertices," in *IEEE Nuclear Science Symposium Conference Record 2006*, vol. 1, p. 104, IEEE, 2006.
121. W. Waltenberger, W. Mitaroff, and F. Moser, "RAVE — a Detector-independent vertex reconstruction toolkit," *Nucl. Instrum. Meth.*, vol. A581, p. 549, 2007.
122. G. Patrick and B. Schorr, "Vertex Fitting of Several Helices in Space," *Nucl. Instrum. Meth.*, vol. A241, p. 132, 1985.
123. P. Billoir and S. Qian, "Fast vertex fitting with a local parametrization of tracks," *Nucl. Instrum. Meth.*, vol. A311, p. 139, 1992.
124. P. Billoir and S. Qian, "Erratum to Fast vertex fitting with a local parametrization of tracks," *Nucl. Instrum. Meth.*, vol. A350, p. 624, 1994.
125. D. Bates and D. Watts, *Nonlinear regression analysis and its applications*. Wiley & Sons, 1988.
126. V. Karimäki, "Effective Vertex Fitting," Tech. Rep. CMS-NOTE-1997-051, CERN, 1997.
127. E. Calligarich, R. Dolfini, M. Genoni, and A. Rotondi, "A Fast algorithm for vertex estimation," *Nucl. Instrum. Meth.*, vol. A311, p. 151, 1992.
128. F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.
129. P. Rousseeuw and A. Leroy, *Robust regression and outlier detection*. John Wiley & sons, 2005.
130. R. Frühwirth, P. Kubinec, W. Mitaroff, and M. Regler, "Vertex reconstruction and track bundling at the LEP collider using robust algorithms," *Comput. Phys. Commun.*, vol. 96, p. 189, 1996.
131. P. Huber and E. Ronchetti, *Robust statistics*. Wiley & Sons, 2 ed., 2009.
132. G. Agakichiev *et al.*, "A new robust fitting algorithm for vertex reconstruction in the CERES experiment," *Nucl. Instrum. Meth.*, vol. A394, p. 225, 1997.

133. K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, p. 2210, 1998.
134. T. Speer, R. Frühwirth, P. Vanlaer, and W. Waltenberger, "Robust vertex fitters," *Nucl. Instrum. Meth.*, vol. A566, p. 149, 2006.
135. W. Waltenberger, R. Frühwirth, and P. Vanlaer, "Adaptive vertex fitting," *J. Phys.*, vol. G34, p. N343, 2007.
136. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, vol. 39, no. 1, p. 1, 1977.
137. R. Frühwirth and W. Waltenberger, "Redescending M-estimators and deterministic annealing," *Austrian J. Statistics*, vol. 37, no. 3&4, p. 301, 2008.
138. J. D'Hondt, R. Frühwirth, P. Vanlaer, and W. Waltenberger, "Sensitivity of robust vertex fitting algorithms," *IEEE Transactions on Nuclear Science*, vol. 51, no. 5, p. 2037, 2004.
139. R. Frühwirth and W. Waltenberger, "Adaptive Multi-vertex fitting," Tech. Rep. CMS CR 2004/062, CERN, 2004.
140. T. Speer and R. Frühwirth, "A Gaussian-sum filter for vertex reconstruction," *Comput. Phys. Commun.*, vol. 174, p. 935, 2006.
141. S. Lang, *Calculus of several variables*. Springer Science & Business Media, 2012.
142. P. Avery, "Applied fitting theory VI: Formulas for kinematic fitting," Tech. Rep. CLEO Note CBX98-37, Cornell University, 1999.
143. ALICE Collaboration, "The ALICE experiment at the CERN LHC," *JINST*, vol. 3, no. 08, p. S08002, 2008.
144. B. Abelev *et al.*, "Performance of the ALICE Experiment at the CERN LHC," *Int. J. Mod. Phys.*, vol. A29, p. 1430044, 2014.
145. G. Aad *et al.*, "The ATLAS Experiment at the CERN Large Hadron Collider," *JINST*, vol. 3, p. S08003, 2008.
146. T. Cornelissen *et al.*, "Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)," Tech. Rep. ATL-SOFT-PUB-2007-007, CERN, Geneva, 2007.
147. ATLAS Collaboration, "Performance of the ATLAS Track Reconstruction Algorithms in Dense Environments in LHC Run 2," *Eur. Phys. J.*, vol. C77, no. 10, p. 673, 2017.
148. G. Aad *et al.*, "Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ =13 TeV," *Eur. Phys. J.*, vol. C76, no. 5, p. 292, 2016.
149. H. Oide, "Improvements to ATLAS track reconstruction for Run-2," *PoS*, vol. EPS-HEP2015, p. 287, 2015.
150. K. Potamianos, "The upgraded Pixel detector and the commissioning of the Inner Detector tracking of the ATLAS experiment for Run-2 at the Large Hadron Collider," *PoS*, vol. EPS-HEP2015, p. 261, 2015.
151. ATLAS Collaboration, "The Optimization of ATLAS Track Reconstruction in Dense Environments," Tech. Rep. ATL-PHYS-PUB-2015-006, CERN, Geneva, Mar 2015.
152. G. Aad *et al.*, "Charged-particle distributions in $\sqrt{s}$ = 13 TeV pp interactions measured with the ATLAS detector at the LHC," *Phys. Lett.*, vol. B758, pp. 67–88, 2016.
153. S. Chatrchyan *et al.*, "The CMS Experiment at the CERN LHC," *JINST*, vol. 3, p. S08004, 2008.
154. CMS Collaboration, "The CMS tracker system project: Technical Design Report," Tech. Rep. CMS-TDR-005, CERN, 1997.
155. E. Brondolin, *Track reconstruction in the CMS experiment for the High Luminosity LHC*. PhD thesis, Technische Universität Wien, 2018.
156. S. Chatrchyan *et al.*, "Description and performance of track and primary-vertex reconstruction with the CMS tracker," *JINST*, vol. 9, no. 10, p. P10009, 2014.
157. CMS Collaboration, "2017 tracking performance plots," Tech. Rep. CMS-DP-2017-015, CERN, 2017.
158. F. Pantaleo, *New Track Seeding Techniques for the CMS Experiment*. PhD thesis, Universität Hamburg, 2017.
159. A. Augusto Alves *et al.*, "The LHCb Detector at the LHC," *JINST*, vol. 3, p. S08005, 2008.

160. R. Aaij *et al.*, "LHCb Detector Performance," *Int. J. Mod. Phys.*, vol. A30, no. 07, p. 1530022, 2015.
161. LHCb Collaboration, "Measurement of the track reconstruction efficiency at LHCb," *JINST*, vol. 10, no. 02, p. P02007, 2015.
162. LHCb Collaboration. Published online at https://twiki.cern.ch/twiki/pub/LHCb/ConferencePlots/TrackEffPLong2015_2012.pdf, 2017.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 14
Distributed Computing

Manuel Delfino

Distributed computing is an established discipline in computer science and engineering. It has evolved over the past 40 years to be one of the most important methodologies for implementing the data processing services needed by almost every activity in society. Distributed computing is still evolving rather rapidly, with major innovations introduced every few years. The aim of this chapter is to introduce the reader to the basic concepts of distributed computing in the context of particle physics, allowing a better understanding of what is behind the scenes when using distributed systems, and providing starting points in order to seek further information on the subject.

14.1 Usage by Particle Physics

Distributed computing, i.e. the coherent use of many computers to accomplish a given task, is extensively used in particle physics. It is particularly suited to simulating and analyzing data from particle collisions, where processing the data of one collision is largely independent from processing data of all other collisions and hence very little communication between the different computers is necessary. This type of distributed computing is called "High Throughput Computing" (HTC),[1] because there is a very large number of quasi-independent tasks to accomplish,

[1]Livny, Miron, et al. "Mechanisms for high throughput computing." SPEEDUP journal 11.1 (1997): 36–40.

M. Delfino (✉)
Departament de Física, Facultat de Ciències, Universitat Autònoma de Barcelona, Barcelona, Spain
e-mail: manuel.delfino@uab.cat

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_14

and the performance perceived by the user is the rate of completion of tasks, or "throughput". This type of computing is complimentary to "High Performance Computing" (HPC), i.e. the execution of a single task at maximum speed on a classical supercomputer. The particular case of event or collision processing requiring little inter-processor communication is an example of "loosely coupled parallel computing" or even "embarrassingly parallel computing".[2,3] Certain kinds of accelerator simulations, particularly those that use ray-tracing techniques, can also be executed as distributed computing tasks. Multiple computer systems can also be interconnected in "tightly-bound" configurations via low-latency networks,[4] essentially yielding a supercomputer. Tightly-bound systems are used in particle physics, for example, for solving equations from theories using numerical methods (e.g. lattice-gauge theories); they are mentioned here for completeness and will not be covered further.

From the decade of the 2010s, the boundary between HTC and HPC has started to blur, for a number of disparate reasons:

Multi-core Applications Detectors have become more complex and have more channels. This impacts the memory footprint of reconstruction, simulation and analysis programs, which has grown substantially. In parallel, processors have become multi-core[5] with shared RAM.[6] The growth in number of cores per processor has been faster than the drop in price of RAM chips, resulting in an effective memory shortage. This means that running an independent copy of the operating system and the application in each core is not economical. The way out is to implement the safe execution of parallel threads[7] of a single copy of the program on each multi-core processor. This introduces a number of dependencies, for example competition for RAM and for input–output services, that break the loose coupling and make the applications behave more like HPC programs.

[2]Wilkinson, Barry, and Michael Allen. Parallel programming. Pearson India, 2004.

[3]Birrittella, Mark S., et al. "Intel® Omni-path architecture: Enabling scalable, high performance fabrics." High-Performance Interconnects (HOTI), 2015 IEEE 23rd Annual Symposium on. IEEE, 2015.

[4]Pfister, Gregory F. "An introduction to the infiniband architecture." High Performance Mass Storage and Parallel I/O 42 (2001): 617–632.

[5]A core corresponds to the Central Processing Unit (CPU) of a classical computer, which is able to execute a single stream of instructions. Technology allows a growing number of ever smaller transistors to be placed on a single chip. However, profiting from these large number of transistors in a single core design would require prohibitively complex processor designs and impossible to achieve clock speeds. The alternative is to populate the chips with many copies of the same processor core, which work independently except for sharing external connections, leading to multi-core processors. Gepner, Pawel, and Michal Filip Kowalik. "Multi-core processors: New way to achieve high system performance." Parallel Computing in Electrical Engineering, 2006. PAR ELEC 2006. International Symposium on. IEEE, 2006.

[6]Random Access Memory, the external solid state memory used by a processor.

[7]A thread is a stream of processing instructions coming from a shared program image which has its own private data instances in processor hardware registers, instruction and data stack and caches, and RAM.

Supercomputers as High Core-Count Clusters The original design of supercomputers was based on a rather small number of the fastest single core processors available at the time, interconnected with custom very low latency links. In addition, the design usually included a much larger amount of RAM than in standard computers. The design of supercomputers has changed completely over the last 20 years.[8] The main driver of this change is the difficulty in building processors with ever shorter clock cycles. The so-called "clock speed wall" has been hit, and practically all processors operate within a narrow range around a few GHz. Hence, the only way to make a supercomputer faster is through parallelism. Modern supercomputers are in fact huge clusters of relatively standard multi-core processors. Low latency links continue to be used and their hardware costs have become much lower. In addition, "latency hiding" techniques sometimes allow the use of standard networks, such as Ethernet. Finally, the amount of RAM per core does not differ much from standard computers. Hence, executing an HTC workload on a supercomputer targeted for HPC is no longer wasteful. At worst, only the low latency interconnect will be underutilized.

Increase of Workloads with Low Input–Output Particle physics has traditionally been a heavy user of HTC systems because they were the least expensive architecture for executing detector track reconstruction and analysis, applications with a relatively low CPU to input–output ratio. The needs for CPU for reconstruction and analysis have grown, however, as detectors have become much more complex. In parallel, the precision needed in the simulations has vastly increased. These high CPU, low input–output applications currently represent the largest computing demand of a modern particle physics detector. Hence, the global workload profile has moved closer to HPC in the last decades.

14.2 Functional Decomposition of a Distributed Computing Environment

It is useful to introduce a functional decomposition, or reference framework, to discuss distributed computing systems. The decomposition that has dominated particle physics data processing since the 1990s, is the "SHIFT Model", introduced in the early 1990s by Robertson and collaborators.[9] It will be described here with some updates of terminology. The basic assumption in the SHIFT framework is that a Local Area Network (LAN) can be built with enough capacity and flexibility so that the rest of the elements in the distributed environment can

[8] Xie, Xianghui, et al. "Evolution of supercomputers." Frontiers of Computer Science in China 4.4 (2010): 428–436.

[9] J.P. Baud, et al. "SHIFT, the Scalable Heterogeneous Integrated Facility for HEP Computing", Proc. Conference on Computing in the High Energy Physics, CHEP91, Tsukuba, Japan, Universal Academic Press.

communicate freely amongst themselves. Furthermore, the environment is loosely coupled and is therefore not sensitive to relatively long round-trip time for network messaging (large network latency). Connected to the network we have the following elements:

- *CPU servers*: Elements that receive input data, perform calculations and produce output data. They do not implement any permanent storage, although they often provide volatile disk storage used to temporarily store data while executing a task.
- *Disk servers*: Elements that store data in a stable and reliable manner with an access latency (defined as time to open a file and receive the first byte of data) which is relatively low. They provide inputs to and receive outputs from the CPU servers. They can also send and receive data to Tape server elements.
- *Tape servers*: Elements that store data in a stable and reliable manner with an access latency which is relatively high. They send and receive data to Disk server elements.
- *Information servers*: Elements that maintain data in a stable and reliable manner about the status of the various elements of the distributed computing environment.
- *Control servers*: Elements that issue commands to trigger operations in other types of servers and coherently update the relevant Information servers.
- *Remote Data servers*: Elements that send and receive data from other security domains, often to Disk server elements via Wide-Area Network (WAN) connections.

The elements of the functional decomposition presented, illustrated in Fig. 14.1, together with the aforementioned powerful network and the maintenance of coherent security[10] domains, constitute a framework which is sufficient to analyze the most commonly used distributed computing environments.

The computing industry places a lot of emphasis on how storage is attached to the computers: Direct-Attached Storage (DAS) does not use a network, attaching storage hardware via an internal communication bus of a computer; Storage Area Network (SAN) uses a short-distance network (originally Fiberchannel, with an increasing tendency towards Ethernet) to couple storage hardware to computers; and Network Attached Storage (NAS) which uses computers hosting storage devices and connected to a LAN (or WAN configured to behave like a LAN) to present storage services to the other computers on the network. These distinctions are important in environments where few computers are used. In the case of particle physics, DAS and SAN are used to construct disk servers, which in turn are exposed almost exclusively in NAS mode.

[10]A more precise term is "Authentication and Authorization Infrastructure", abbreviated as AAI or AA.

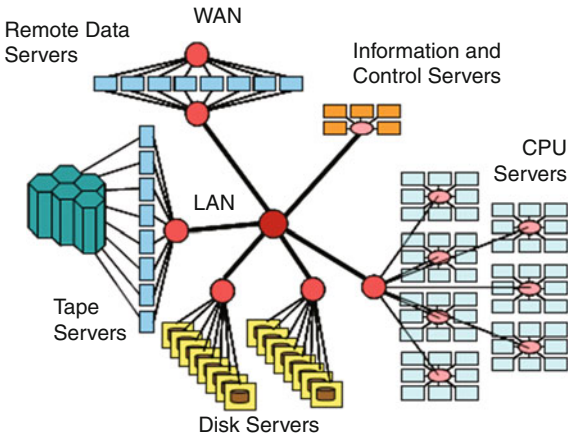


Fig. 14.1 Functional decomposition of a distributed computing environment

14.3 Data Processing Clusters

The most common way to deploy a distributed computing environment is a "cluster".[11] A cluster consists of a number of CPU, Disk and Control servers connected on a LAN[12] and sharing status information through local Information servers while maintaining a *single, coherent, security domain*. A cluster may also implement Tape servers and Remote Data servers.

Historically, the first complete and reliable commercial implementation of a cluster was the VAXcluster[13] (or VMScluster), implemented by Digital Equipment Corp. starting in 1983 on VAX computers and workstations using the VMS operating system. Today, essentially all clusters used by particle physics are implemented using the Linux operating system[14] as a basis, and adding a number of additional packages for handling authentication and storage.

[11] Unfortunately, after many decades of using the term "cluster" as in the present text, some commercial firms have taken to using the terms "Grid" or "Cloud" to mean "cluster". This is considered by the author as incorrect and confusing.

[12] There are also implementations using several LANs transparently interconnected via WANs, but maintaining a single security domain. This configuration is often used in business environments to gain fault-tolerance and disaster recovery. Problems related to performance limitations of the WAN, as well as inter-institutional security management complications, limit the use of this configuration in particle physics data processing. Interconnection of clusters via a Grid or a federated Cloud is used instead.

[13] Kronenberg, Nancy P., Henry M. Levy, and William D. Strecker. "VAXcluster: a closely-coupled distributed system." ACM Transactions on Computer Systems (TOCS) 4.2 (1986): 130–146.

[14] Linux (or more precisely GNU/Linux) is mentioned specifically as it has become the dominant Unix-like operating system, and has become better known than the original Unix operating system. Most of the remarks in this chapter, however, are equally applicable to other Unix-like systems, notably macOS and FreeBSD.

14.3.1 Authentication and User Identification

The machinery behind authentication and user identification is often poorly understood by users, which leads to usage and security problems. It is important to gain an understanding of the basic concepts, especially as more complex distributed environments, such as Grids and Clouds (see sections below), have come into widespread use.

A number of Information servers are deployed to integrate individual computers to form a cluster.[15] The most crucial ones are the ones that generate a single, coherent, security domain.[16] Security under Linux and most other operating systems is based on the concept of an "account" which, ideally, is used exclusively by a single trusted individual. An account is represented externally by an alphanumeric string (the "username") and internally by a numeric code (the *userid* or *uid*). Although modern operating systems allow very long usernames, local restrictions may apply in order to maintain compatibility with older system software and utilities (the most common restriction is a maximum length of eight characters). Each account is associated to specific directives (called "rights" or "privileges") to allow or disallow access to operating system services, and can also be used to control access to files and other resources directly or via Access Control Lists ("ACL"). Since people often work in teams, it is useful to associate an account with a "group" named by an alphanumeric string and represented internally by a numeric code (the *groupid* or *gid*). Certain rights and, very importantly, file access control can thus be quickly managed for all accounts belonging to a group.

The process of authentication involves an exchange of credentials which establishes the identity of the user wanting to access a system which results in the creation of a process running under the corresponding *uid*. Authentication has been traditionally accomplished by a "password", an alphanumeric string which, ideally, is known only to the individual owner of an account. Password related issues are major contributors to computer security problems. Practically all computers are nowadays connected directly or indirectly to the Internet. An inherent weakness of a cluster (which is more than compensated by the gained functionality) is that having a single security domain means that gaining access to any component of the cluster grants access to all of it. The components of clusters directly connected to the Internet are under constant attacks, the most common being automated attempts to guess passwords (usernames are easily obtained from public information sources, such as Web pages listing email addresses). Attacks from within the cluster must also be considered, especially in clusters with many accounts. A

[15]The term Information server is used in a broad manner. For example, Domain Name Service (DNS) servers which translate alphabetic Internet addresses to numerical, and Network Time servers which ensure all cluster elements have their clocks synchronized, are considered Information servers.

[16]The actual implementation is a highly technical matter, using NIS, ldap, Microsoft Active Directory or other secure database sharing schemes.

password which is easily guessed is denominated "weak". Simple measures exist to avoid weak passwords:[17] The password should be as long as possible and contain a mix of numbers and upper and lower case letters. Many installations require users to change their passwords periodically in order to improve computer security. A very dangerous practice which should be completely avoided, but which is unfortunately common practice in the particle physics community, is the "sharing" of personal accounts (and hence, their passwords) or the use of "service" accounts with passwords known by dozens of persons, written in documents and blackboards or even posted on Web sites. Modern operating system features make these practices completely unnecessary, as they can be configured so that each user can first authenticate with their personal account and then gain access to a common environment to perform the tasks required.

Portable devices ("laptops" and "smartphones") can be deployed so that they are elements of a cluster and therefore part of the security domain. Since the portable computer must remain useable when temporarily disconnected from the network, the operating system will keep a local copy (or "cache"[18]) of the security data, including the passwords. This means that a stolen portable computer is a computer security threat. Care should be taken in configuring the "suspended" or "hibernated" modes in portable computers to ask for a password when they are turned back on.

Another major issue with authentication based on usernames and passwords is that each cluster (and each online service such as electronic mail, social networks, and document and file repositories), being a separate security environment, requires a separate username and password and has its own policies for requiring password changes. Users are becoming overwhelmed in keeping track of their usernames and passwords and their reaction is often to use weak passwords or keep a list of passwords in a file which can be stolen. In order to reduce username/password proliferation, many organizations are linking their clusters and online services to shared authentication servers, thus providing a "single-sign-on" environment.

In the 2000s, organizations, most notably the CERN Large Hadron Collider experiments, started to use an alternative authentication and authorization frame-

[17]Users should consult their local security rules for specific recommendations on password choices.

[18]The term "cache" is used in computer architecture to describe a local copy of a limited amount of data which is normally stored elsewhere. The introduction of this local copy serves to increase the performance and efficiency of use of the element connected to it, by smoothing out peaks and valleys in storing and retrieving information, or avoiding retrieves altogether if the same information is needed repetitively. The cache is transparent to the element accessing the information; if the data needed is in cache, it will be delivered from there (a cache "hit") whereas if the data is not in cache, the normal access path to the normal storage place will be used (a cache "miss"). The key to successful implementation of a cache is the right choice of algorithm to choose which data to copy and keep in the cache. Examples are high-speed memory caches in CPU chips which hold local copies of instructions and data, and memory caches in disk controllers which hold local copies of disk blocks.

work based on digital certificates.[19] This was one of the pillars for constructing the Worldwide LHC Computing Grid (WLCG), later generalized to the European Grid Infrastructure (EGI), the U.S. Open Science Grid (OSG) and others. The certificates used are based on the X.509 standard, supported by all major Web servers and browsers. These schemes were very successful and are still being used, allowing tens of thousands of computers to provide coherent services to thousands of users. They are, however, extremely difficult to manage and expensive to operate. Hence, the tendency is to abandon these schemes in user-facing services, keeping them mostly for machine-to-machine services.

Industry and academia are putting a lot of effort on the emerging area of "distributed" authentication and authorization (AA) schemes.[20] The ultimate aim is to simplify the AA process and make it easier to use, enabling "single-sign-on" to a wide variety of resources, a functionality that users have come to expect from their experience with social networks. Unfortunately, progress has been rather slow and solutions proposed by industry and academia are often incompatible. Part of the problem is that there are two competing standards: SAML[21] and OpenID.[22]

The main problem, however, is that the usage by large collaborations, such as those in particle physics, ideally requires the simultaneous use of many different AA sources (something known as Identity Federation). The largest Identity Federation currently deployed is the *eduGAIN* federation,[23] where the AA sources come from the universities and research institutes who employ the users. *eduGAIN* is modeled after the successful *eduroam* system[24] used to grant worldwide access to academic WiFi networks worldwide. Solving the general access problem is much more difficult, however, as it requires the maintenance of many more attributes for each user. For example, each university and research institute would have to include and maintain in their databases which experiment or project each employee is participating in, something which is not practical. In order to solve this issue, hybrid schemes are being worked on, where the authentication would come from *eduGAIN* but the authorization information would come from an attribute server managed and operated by a specific experiment or project. This still leaves the problem of reliably operating a service build from thousands of independently managed AA servers with varying degrees of service quality.

[19]Thompson, Mary R., Abdelilah Essiari, and Srilekha Mudumbai. "Certificate-based authorization policy in a PKI environment." ACM Transactions on Information and System Security (TISSEC) 6.4 (2003): 566–588.

[20]These schemes centralize AA information and then make it available to a distributed set of heterogeneous resources, hence the term "distributed".

[21]Rosenberg, Jonathan B., and David L. Remy. Securing web services with WS-security: Demystifying WS-security, WS-policy, SAML, XML signature, and XML encryption. Sams, 2004.

[22]https://openid.net/developers/specs/

[23]López, D. "eduGAIN: Federation interoperation by design." TERENA Networking Conference. 2006.

[24]López, Gabriel, et al. "A proposal for extending the eduroam infrastructure with authorization mechanisms." Computer Standards & Interfaces 30.6 (2008): 418–423.

Some interesting alternatives are starting to emerge, which avoid Identity Federation. One approach is to identify a single reliable provider for AA, for example the OpenID service from ORCID[25] combined with an appropriate attribute server. Another approach is for a project or experiment to deploy a single AA scheme available to services *for that project* around the world. This can be accomplished in a simple and economical manner by re-using the project's personnel database and exposing it through the Internet in a secure manner using an *ldap* server.[26] These type of implementations of single-sign-on are ideal for service providers that serve a single project. On the other hand, they do shift the diversity problem to the service providers, which must locally adjust crucial attributes such as *uid* and *gid* in order to avoid duplications in multi-project or multi-experiment environments.

14.3.2 Processing and Storage

Two configurations of CPU servers are usually deployed:

- "Batch workers", which execute tasks (jobs) that don't require user interaction and are scheduled using a "Batch" system[27] or a more sophisticated resource harvester, such as HTCondor.[28]
- "Interactive nodes", where users can connect (or "log-in") and perform work that requires interaction via a screen, keyboard and mouse.

Several configurations of Disk servers are usually deployed:

- "Network File servers", based on industry-standard or widely used protocols, such as NFS,[29] smb2[30] or http with WebDAV,[31] are used to provide "home" and "project" directories holding text and binary files of relatively small size. These directories appear in the CPU servers as "mount points" or "network shares" that largely behave as a local, conventional disk resource. The POSIX input–output

[25] Haak, Laurel L., et al. "ORCID: a system to uniquely identify researchers." Learned Publishing 25.4 (2012): 259–264.

[26] Johner, Heinz, et al. Understanding LDAP. Vol. 6. IBM, 1998.

[27] See for example Yoo, Andy B., Morris A. Jette, and Mark Grondona. "Slurm: Simple linux utility for resource management." Workshop on Job Scheduling Strategies for Parallel Processing. Springer, Berlin, Heidelberg, 2003.

[28] Fajardo, E. M., et al. "How much higher can HTCondor fly?" Journal of Physics: Conference Series. Vol. 664. No. 6. IOP Publishing, 2015.

[29] Shepler, Spencer, et al. Network file system (NFS) version 4 protocol. No. RFC 3530. 2003.

[30] French, Steven M., and Samba Team. "A New Network File System is Born: Comparison of SMB2, CIFS and NFS." Linux Symposium. sn, 2007.

[31] Goland, Yaron, et al. HTTP Extensions for Distributed Authoring – WEBDAV. No. RFC 2518. 1999.

standard[32] is the most often used definition of this behavior. The implementations are far from trivial, but fortunately they are widely used and hence available as packages fully supported by operating systems, or even as part of the operating system. These protocols support operations such as simultaneous access by multiple users in read/write/append mode as well as record-level locking. This forces the deployment in CPU servers of some Control and Information server functionality, with information being constantly exchanged between all nodes accessing a mount point in order to keep file system coherency and distribute lock information. In order to fulfill performance expectations, these Control and Information functionalities must often be implemented as drivers, often executing inside the operating system kernel. The price of this is the risk of all CPU servers becoming inoperative (or "hung") if a network file server has a fault or the status information becomes corrupted. Some Network File servers, such as the CernVM-FS[33] system used for software installation, serve very specialized purposes. They implement only part of the functionalities, for example serving files only in read-only mode, with the benefit of having a smaller impact on the operating system environment.

- "Database servers", are disk servers[34] which organize their data using database products such as Oracle, MySQL or PostgreSQL. They are mostly used to store and manage detector configuration and calibration data, though on occasion they can store high-level analysis data such as the "Event Tags" used by some experiments.[35] CPU servers read these data through the network by issuing database queries and retrieving the results. Database servers must be configured in accordance to the type of most-often used queries (a process called "tuning") in order to achieve the required performance. Because the overall cost of purchasing and maintaining a Database server is high, situations often arise where the deployed capacity is insufficient to serve all the CPU servers unless specific caches or buffers[36] are deployed. An emerging alternative type of database

[32]Gallmeister, Bill. POSIX. 4 Programmers Guide: Programming for the real world. "O'Reilly Media, Inc.", 1995.

[33]Blomer, Jakob, et al. "Status and future perspectives of CernVM-FS." Journal of Physics: Conference Series. Vol. 396. No. 5. IOP Publishing, 2012.

[34]It may seem odd to classify a database server as a Disk server. It is considered correct in this context, as their usage pattern is unusual in that the record update frequency is very low compared to the record reading frequency. Note that the technical database servers used to implement Data and Tape server features are considered Information servers.

[35]See for example Cranshaw, Jack, et al. "Event selection services in ATLAS." Journal of Physics: Conference Series. Vol. 219. No. 4. IOP Publishing, 2010.

[36]The term "buffer" is used in computer architecture to mean "temporary storage location". Buffers are used to accumulate information and then transmit it as a block, or to receive information as a block and then distribute it. The difference between a buffer and a cache is that in the case of buffers all the information is temporarily stored, whereas a cache implements an algorithm to store only the most relevant information. In addition, whereas a cache is always transparent, a buffer may be transparent or it may allow explicit buffer manipulation by the application.

servers are those based on "Big Data" tools, such as NoSQL databases[37] with underlying *Hadoop* storage management systems.[38]

- "Data servers", based on products such as dCache,[39] GPFS,[40] Lustre,[41] EOS[42] and DPM,[43] are used to hold the large-size binary files which contain physics data. They use a set of specific Control and Information Servers to present to the other elements of the cluster, in particular to the CPU servers, an interface approximating a conventional (POSIX-compliant) disk resource, but with some important differences in order to achieve the large capacities and high performance needed for particle physics applications (and data-intensive applications in other fields). Much of the optimization is possible because of the particular usage pattern of particle physics data, which is essentially written once and then read many times.[44] The key point is to separate[45] the "namespace" information (directory tree or file folder structures, file attributes such as name, creation date, etc.), the "file storage" information (physical location of the blocks that comprise a file, whether it is open for read, write or append, etc.) and the "file data" information (the actual data blocks that comprise the file). This separation makes it possible to serve the data blocks from a large number of Data servers with a combined capacity far exceeding what can be provided by a conventional file system. In addition, the data blocks of a given file may be stored in multiple servers in order to enhance read performance or provide high-availability (for reading).[46] The cost of this separation (apart from the complexity of the multiple

[37] Han, Jing, et al. "Survey on NoSQL database." Pervasive computing and applications (ICPCA), 2011 6th international conference on. IEEE, 2011.

[38] Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. Ieee, 2010.

[39] Millar, Paul, et al. "Storage for advanced scientific use-cases and beyond." Parallel, Distributed and Network-based Processing (PDP), 2018 26th Euromicro International Conference on. IEEE, 2018.

[40] Schmuck, Frank B., and Roger L. Haskin. "GPFS: A Shared-Disk File System for Large Computing Clusters." FAST. Vol. 2. No. 19. 2002.

[41] Schwan, Philip. "Lustre: Building a file system for 1000-node clusters." Proceedings of the 2003 Linux symposium. Vol. 2003. 2003.

[42] Peters, A. J., E. A. Sindrilaru, and G. Adde. "EOS as the present and future solution for data storage at CERN." Journal of Physics: Conference Series. Vol. 664. No. 4. IOP Publishing, 2015.

[43] Alvarez, Alejandro, et al. "DPM: future proof storage." Journal of Physics: Conference Series. Vol. 396. No. 3. IOP Publishing, 2012.

[44] Aside from the write-once/read-many access pattern, particle physics datasets almost never have data appended to them nor have their already existing records updated. For an extreme contrast, consider the access pattern of bank account datasets, which constantly have data appended and updated.

[45] Conventional file systems, whether local or networked, tightly bind the management of the three types of information described in order to optimize file open/close operations and implement sharing of files in read/write/append mode.

[46] This is conceptually similar to the "striping" of files across many disks performed by RAID controllers on DAS devices.

Information Servers) is a relatively high overhead and latency for file open/close operations and the loss of full POSIX compliance. In some cases, applications have to use input–output methods specific to the data server product, such as *Xrootd*,[47] whereas in other cases standard protocols, such as NFS or http or subsets of POSIX I/O can be used. Historically, namespaces have been confined to single clusters. There are, however, some interesting implementations of global namespaces, using for example a hierarchy of distributed name servers coupled through redirection[48] techniques offered by various protocols such as *Xrootd*, WebDAV or http.[49] Many commercial systems use similar implementations, but with different design criteria, for example the requirement of serving hundreds of thousands of users at relatively low performance in a social network.

Tape servers are hidden from the user,[50] except in very special cases such as data recording of experimental data. Nevertheless, users and system managers should be vigilant, as the user may generate access patterns which make extremely inefficient use of the underlying tape cartridge system. Tape servers and their related tape cartridge systems are quite complex technologically. Exploring this is beyond the scope of this chapter, but a few remarks are appropriate. The main motivation for introducing Tape servers is that they provide storage which is less expensive than Disk servers, with additional features such as guaranteed read-only access, protection against erroneous deletion by users, lower probability of data loss due to hardware failure and lower power usage. As their name implies, tape servers have been historically implemented using storage on magnetic tape cartridges. Implementations for particle physics are of the "Active Tape" nature, using specific software to manage the tape in a much more agile manner than traditional tape backup systems, as well as providing file by file access. Different computer centers use different software packages for Active Tape management. Examples are HPSS[51] and Tivoli Storage Manager,[52] both developed by IBM, Enstore[53] developed by Fermilab and Castor2[54] developed by CERN. Decreasing prices for commodity disks, however, are driving an emerging area where Tape

[47]Dorigo, Alvise, et al. "XROOTD-A Highly scalable architecture for data access." WSEAS Transactions on Computers 1.4.3 (2005).

[48]For example, redirection is a standard feature in the http protocol used for the World Wide Web.

[49]See for example Bloom, Kenneth, and Cms Collaboration. "Cms use of a data federation." Journal of Physics: Conference Series. Vol. 513. No. 4. IOP Publishing, 2014.

[50]Conventional backup to tape of files in home and project directories is usually performed as part of the general operation of clusters using standard commercial or open-source packages. This is an important function but, being completely conventional, is not considered further in this context.

[51]Teaff, Danny, Dick Watson, and Bob Coyne. "The architecture of the high performance storage system (hpss)." (1998).

[52]IBM Corporation, Tivoli Storage Manager; see http://www.ibm.com/software/tivoli/products/storage-mgr/

[53]Bakken, Jon, et al. "Enstore Technical Design Document." Fermilab-JP0026 (1999).

[54]Presti, Giuseppe Lo, et al. "CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN." MSST. Vol. 7. 2007.

servers are implemented on inexpensive commodity computers with large disks. This is in turn stimulating multi-layer implementations, for example the D2D2T (disk-to-disk-to-tape) scheme, where Disk servers store and retrieve data on disk-based pseudo-tape servers, which in turn store and retrieve data on "real" magnetic cartridge Tape servers. Both disk-based tape servers and D2D2T schemes are important because the large growth in data volumes in particle physics is the result of a very modest growth in file size multiplied by a large growth in the number of files, resulting in a situation which can be very inefficient for traditional tape systems (the so-called small file problem in tape cartridge storage).

A "Hierarchical Storage Manager" (HSM) is implemented between the Data servers and Tape servers using specific Control and Information servers. HSM systems are not a common feature of conventional computing environments. HSM systems that fulfill the needs of particle physics data processing are even less common. Furthermore, the interaction between Data servers and Tape servers can be rather complex and no clean, practical interfaces are available. Therefore, each data center implements its own customized solution using a combination of community-written and commercial software. Examples of Data server/HSM implementations are dCache interfaced to Tivoli or Enstore; GPFS interfaced to IBM HPSS; and CERN's Castor-2 which has its own data server and HSM implementations. Depending on the solution, the HSM system may be more or less intertwined with the Data server software, although the tendency is to separate as much as possible the HSM and Data server implementations.

Experiment specific Information servers are deployed in order to manage and make accessible a view of all their files containing custom details about their characteristics, for example trigger conditions, detector configuration or beam type. This is often called a "File Catalog" and may in itself be a quite complex database.

14.4 The Main Workflows of Data Processing in an Experiment

Most experiments implement two "organized" data processing workflows: one for data coming from the actual detector (awkwardly referred to as "real" data) and another for simulations resulting from the techniques described in the contribution "Detector Simulation" (awkwardly referred to as "Monte Carlo" data). An additional very important "organized" workflow derives calibration constants from detector data taken either in normal mode or in special conditions for calibration (cosmic rays, single beam, sending pulses to the front-end electronics, etc.). Altogether, these three workflows are referred to as "production" and constitute the process by which "raw" data is transformed into "analysis datasets". Workflows related to the use of these analysis datasets by individual physicists are described in a later section.

The production workflows involve a number of tasks,[55] such as event reconstruction, simulated event production, data reduction, event filtering, and calibration data processing.

14.4.1 Event Reconstruction

Data arriving from an experiment's data acquisition system is called "raw" data. It is organized in "event records", each record corresponding to a particle interaction "event" which is relevant in the context of the experiment. An event record may or may not correspond to a single collision between particles, though particle physicists often refer to data arising from a single collision as "an event". In a collider, an event record may correspond to data collected from a single beam crossing, which depending on the beam type, luminosity and angular coverage may contain data arising from several particle collisions.[56] In a fixed target experiment an event record may correspond to data collected over a full beam spill. In an astroparticle experiment an event record may correspond to data collected starting from a trigger signal for a certain length of time. Furthermore, due to performance requirements of data acquisition systems, raw data is often compressed, encoded or packed into rather complex data structures, which are often not very convenient for direct use in reconstruction programs. Therefore, a raw event record is usually read from disk and immediately "unpacked" into simpler but larger data structures where the digitized signals are represented by standard integer or floating point numbers.

The output of the most complete simulations (often called "full" simulations, see below) can be stored in "packed raw" format identical to the detector data, or in "unpacked raw" format in which case the unpacking step is omitted in reading simulated event records.[57] Thereafter, simulated event records are treated identically to detector "real" data.

A number of event records are written by the data acquisition system to a raw data file, which is the main data item delivered from the "online" to the "offline" environment. The organization of event records into raw data files is very often unrelated to the experimental conditions and simply reflects technical constraints such as the size of disk buffers in the online and offline clusters.

[55]The names of these tasks are not standardized. Therefore readers will have to identify the corresponding task names for their respective project.

[56]In electron-positron collisions there is a single collision most of the time, but there may be occasionally an annihilation event together with a two-photon collision, for example. In proton–proton collisions there are on average 20 collisions per beam crossing at the nominal luminosity of the original LHC configuration.

[57]Even though it would be more economical to store simulated data in "packed raw" format, this is often not possible. For example, the simulations may be needed before the detector data acquisition system is fully designed and implemented, and therefore the exact format for packed raw data will not be known.

The output of the reconstruction arises from the application of the techniques described in the contribution "Patter Recognition and Reconstruction" and consists of data structures containing tracks, clusters, etc. These data structures are written as event records, which may or may not have a one-to-one correspondence to the event records read on input. For example, the reconstruction program may disentangle data from different collisions contained in a single raw data event record and write the corresponding reconstruction output as separate event records for each collision.

The reconstruction program needs detector configuration and calibration data that often varies with time, though reasonably slowly compared to the event rate.[58] Therefore, the reconstruction program accesses a "conditions database" in order to fetch the configuration and calibration data matching the conditions under which a particular event was acquired. Alternatively, if stable and reliable calibration data are available at the time of acquisition, it may be stored within the raw data files, interspersed with the event records.

Event records are, in principle, all independent from one another and therefore suitable for loosely-coupled parallel processing. Early implementations[59] would process one raw data file at a time, in essence treating the raw data file as a buffer and implementing an "event server" to distribute individual events via the network to CPU servers. This methodology has been largely abandoned for event reconstruction, although it remains a good choice for parallel processing of analysis data sets. Modern event reconstruction clusters or "farms" are configured so that each CPU server executes jobs that process one or more complete data files at a time. The CPU servers read raw and calibration data and write reconstructed data in an autonomous fashion to the Disk and Database servers. Hundreds or even thousands of jobs can run simultaneously in a single cluster.

Care must be taken to handle aspects that may introduce coupling and break the highly parallel processing scheme. Consider, for example, one of the most common pitfalls that breaks parallelism: simultaneous access to the same file by many jobs. For example, the executable image of the reconstruction program is the same for all jobs and may be quite large. Therefore, each job that starts must read the image file from a Network File server. In an environment where thousands of jobs are executed, the number of jobs simultaneously reading the image file may be

[58] There are exceptions to this rule, especially for neutrino beam or non-accelerator events, where calibration data may in fact be updated faster than the event rate.

[59] Event processing in parallel dates back to the 1970s when physicists used their experience in building digital electronic boards for detectors to design inexpensive processor boards, called emulators, which executed the instruction set of the IBM System 370 mainframe computer and interfaced to its input-output channel by emulating a tape drive. Starting in the mid-1980s, parallel event processing was mostly done using specially configured clusters (called "farms") of scientific workstations with Reduced Instruction Set CPUs (RISC) running the VMS operating system from Digital Equipment Corp. and a number of variants of Unix, interconnected by proprietary networks or by daisy-chained Ethernet networks. Starting in the mid-1990s to date, essentially all reconstruction is run on clusters of widely marketed, industry standard "commodity" computers using the x86 instruction set supported by Intel and Advance Micro Devices CPU chips and interconnected by switched Ethernet networks.

high enough to overload the Network File server, therefore creating a performance bottleneck.[60] Caching of the image file on a local disk of each CPU server can be used to restore parallelism more economically compared to deploying a more powerful Network File server. A similar situation arises from the access by all jobs to a unique conditions database. Parallelism is harder to restore in this case and requires tuning of the database server and the caching algorithm, or deployment of very powerful (and expensive) replicated database servers.

Raw data files must often be retrieved from tape, and parallelism may be broken from the fact that a single tape cartridge is used to store multiple raw data files. Many HSM systems are unable to handle multiple, asynchronous, closely spaced requests for files on the same tape. They queue the requests and mount and dismount the tape cartridge in order to read each of the files, a slow procedure as it involves mechanical actions. As individual raw file sizes have historically increased more slowly than total raw data volume, the number of files per tape cartridge has increased and parallelism breakage by tape access has become quite common. In order to circumvent these problems, production managers run special jobs to "pre-fetch" in an efficient way lists of specific datasets from tape to Disk servers. Efficient pre-fetching requires detailed knowledge of the HSM software and the contents of the tape cartridges in order to minimize the number of tape mounts.

Most modern experiments strive to run a production workflow for reconstructing the raw data that keeps up with the data acquisition, called "quasi-online" or "prompt" reconstruction. The quality of the calibration constants which are available for this workflow is often limited and therefore the reconstructed output may be of limited value for final physics results; the output is very valuable, however, to monitor detector performance and to derive more definitive calibration constants.

Frequently, reconstruction of a particular raw data file must be repeated a few times, in order to incorporate improvements or correct errors in calibration constants and reconstruction algorithms. This is called "reprocessing" and applies to both detector and simulated data.

14.4.2 Event Simulation

Production of simulated events usually starts with the generation of physics final-states using an "event generator". The CPU time needed for this step is usually

[60]To understand how parallelism is broken, consider that the Network File server handles read requests in parallel up to a certain limit and will serialize into a queue all requests exceeding this limit, breaking parallelism. To understand how coupling is introduced, consider that the event records are coupled through the fact that they are all being reconstructed using the same reconstruction program image file. This can be exacerbated by additional coupling at the hardware level, for example when many cores compete for the same Ethernet network interface in modern multi-core computers.

small, as is the size of the output file (which is often called a "four-vector file").[61] A single (non-parallel) job can usually produce enough four-vectors for a whole simulation campaign. The situation is completely different in the next step, where the four-vectors go through a full detector simulation which is very CPU time consuming. Many jobs must be run in parallel reading four-vectors and writing simulated event records[62] to Disk servers. The data rate per job is not very high, but the integrated output rate can be substantial since there are many CPU servers running simultaneously, an issue that should be taken into account. Simulation is often run on many separate clusters and the oversight of simulation productions can consume substantial amounts of time from a large number of physicists. Deployment of Grid interfaces on clusters (see below) has simplified these tasks somewhat. A recent tendency is to use "production management systems" based on databases to keep track of simulation jobs. The most tedious part, however, is the detection and correction of errors coming from the execution of the simulation program itself, which currently is not automated in part due to lack of a systematic approach to error signaling.

14.4.3 Data Reduction and Event Filtering

Appropriate data representations for physics analysis are quite different from the raw data format. They can also be quite different from the reconstruction output format for several reasons, some related to enabling limited reprocessing capabilities from the reconstruction output itself, while others are related to signal-to-noise issues. Reconstruction output may have a format optimized for limited reprocessing without access to the raw data, for example recalculating vertices from tracks or track and calorimeter cluster matches. This format may not be optimal for analysis. A simple example can illustrate this: a typical analysis condition is to access all tracks that have a given minimum number of points (or "hits"); hence, the optimum data structure is a list of tracks, each track having the property "number of points". The reconstruction output, however, will most likely be stored as a list of points, each point having a property of belonging to a track, a data structure optimized for reprocessing the assignment of points to tracks, but completely "backwards" from the one needed for optimal analysis. Reconstruction output will often have additional information stored, related to enabling reprocessing, which makes it too bulky for practical use in analysis. It is often organized keeping a

[61] A single generation step is described for simplicity. Note that in some cases the output of an event generator may need to be post-processed by another physics simulation program, for example to turn quarks and gluons into jets.

[62] A single step is described for simplicity. Note that in some cases the output of the detector simulation may have to be post-processed to generate simulated event records, for example by adding simulated pile-up events for proton–proton collisions in the LHC.

one-to-one relationship to raw data files, which in turn may have little to do with physics analysis. Furthermore, it usually has output for all raw data records, which depending on detector triggering and physics signal-to-noise conditions may be very large compared to the optimal for physics analysis.

These considerations lead to the definition of "reduced" or "analysis" datasets along two complementary lines. Data reduction takes the reconstruction output and generates data structures optimized for analysis by removing information related to reprocessing and "turning around" data structures related to tracks, vertices, calorimeter clusters, etc. to be optimal for analysis. Particle identification by correlation between different detector elements is often done (or re-done) during data reduction and, when combined with Event Filtering, may be optimized for particular physics processes (heavy quark identification, rejection of electrons, etc.). The historical name for the output of data reduction is "Data Summary Tape" or "DST" and successively smaller outputs have been called "mini", "micro" or even "nano" DSTs.[63] A nano-DST may simply consist of a list of "reconstructed particles", each stored with a few properties such as particle ID probabilities and momentum components. A nano-DST is in some sense the experimental counterpart to the event generator output.

DST files often do not mirror the raw and reconstructed file organization scheme. At a minimum, DST files contain the DST records that correspond to many raw files, in order to ease the management of large numbers of datasets and avoid handling very small files. In accelerator experiments, DSTs may be organized by running conditions. In fixed-target experiments where on-beam and off-beam data are alternatively taken, separate DSTs for each may be produced. In astroparticle experiments, which often behave like telescopes, DSTs will often be organized by observed source, merging data from many observation days or nights. These re-organizations of data are special cases of event filtering.

Event filtering[64] may be performed before, after or during production of any DST-like output. The purpose is to group together in particular files the DST output most relevant for a particular analysis, thereby avoiding repetitive reading of the rest of the data by analysis programs. Event filtering can be done according to many criteria, but is most often performed as a pre-selection for particular physics analyses, especially in situations of unfavorable signal-to-noise ratio such as the LHC experiments. Output datasets are often referred to by the name of the physics process pre-selected and are written in an identical format as the original DST-like datasets.

[63] The names of Data Reduction output datasets are not standardized. Therefore readers will have to identify the corresponding dataset names for their respective project.

[64] Event Filtering in this context should not be confused with the actions taken by high-level trigger "Event Filters".

14.4.4 Processing of Calibration Data and Calculation of Calibration Constants

Reconstruction of data from detectors requires a large amount of calibration constants, many of them time-varying. Raw calibration data, such as temperatures, pressures and voltages are recorded periodically during data taking. Calibration data related to response and alignment of different detector elements can be obtained by generating artificial events with specialized calibration devices, for example pulsed lasers. Alternatively, they can be calculated from the raw physics data itself, using better understood reactions or additional events occurring in parallel to the physics, such as the passage of cosmic rays.

Data input to calibration procedures often comes from diverse sources, and handling it poses challenges that are often underestimated. For example, laser pulses for calibration may be recorded in parallel to the physics data and stored by the online system interspersed in the raw data files, thereby requiring the offline to access all raw data files in order to process the laser pulses, a rather tedious procedure. Conversely, sometimes calibration data are stored in a *potpourri* of small files and databases of diverse formats, introducing unnecessary complexity in the handling of these data.

There is no standardized framework to handle the output of all calibration procedures, although the recent trend is to centralize it into a single calibration (or "conditions") database that is accessed through heavily cached methods by instances of the reconstruction, DST generation and analysis programs. Care must be taken to serve this database with adequate resources, and to ensure efficient and cached access to it in order not to break parallelism as described earlier.

The flow of raw and derived calibration data and the related data processing tasks should be studied with equal care as that of the physics data, and solutions should be implemented to ensure efficient and agile data access for calibration purposes.

14.5 Interactive Analysis Using Clusters

The last stages of physics analysis are performed by individual users using interactive tools such as ROOT,[65] Jupyter Notebooks[66] or SWAN.[67] They apply

[65] See http://root.cern.ch

[66] See https://jupyter.org/about. Athough Jupyter Notebooks are language agnostic, they are closely associated with the Python language. They are increasingly popular in particle physics and many other scientific disciplines, given the rise of the use of Python by them. Jupyter Notebooks must be combined with suitable data-access modules. For example, the SWAN project at CERN combines Jupyter Notebook with ROOT input–output modules to create an interactive analysis platform.

[67] Danilo Piparo, Enric Tejedor, Pere Mato, Luca Mascetti, Jakub Moscicki, Massimo Lamanna, SWAN: A service for interactive analysis in the cloud, Future Generation Computer Systems, Volume 78, Part 3, 2018, Pages 1071–1078, ISSN 0167-739X, https://doi.org/10.1016/j.future.2016.11.035

selection criteria to events in DST-like files or in personal or group files derived from them and written directly in the format of the analysis tools. The basic record in these files is often a collection of ordered lists of properties, or "n-tuples".

Distributed computing is used for preparing the input files for the interactive tools, using the same methodologies described earlier for running reconstruction and data reduction jobs. In addition, since the application of selection criteria and the generation of histograms from n-tuple files can often be mapped to highly parallel tasks (if each n-tuple record is independent of any other), efforts have been made to provide parallel "back-ends" to interactive analysis tools, for example the PROOF back-end to ROOT or the use of *Hadoop* in conjunction with Jupyter Notebooks. The idea is simple in principle: the user issues a command that specifies some actions to be performed on all records; the back-end automatically generates and executes in parallel a number of tasks reading the data and writing partial result files; when all parallel tasks have finished, the partial results are merged to a global result which is presented to the user via the interactive front-end. Speed is gained, in principle, by using many CPUs and, more importantly, many input data channels in parallel. In addition, since the interactive user pauses to examine the results, resources can be used more efficiently by serving many users from a single parallel back-end. Practical implementation with good performance, however, can be extremely complex, as they involve running a cluster at very large peak input–output rates and clever scheduling of bursts of parallel tasks.

14.6 Multiple Sites and Grid Computing

Grids were defined in the late 1990s as a new manner of integrating services from clusters across multiple sites.[68] Practical development and deployment of Grids was originally led by the particle physics community, especially the LHC collaborations. Grids are convenient for two main purposes: data distribution to many end-users who are geographically dispersed, and data processing load sharing across many computing centers ("resource" centers).

The main idea of a Grid is to abstract the most commonly used interfaces for authentication, access control, job submission and data access into a unique "meta-interface", which is the only interface exposed to end-users. Therefore, multiple sites are integrated for the end-user under the illusion of a single "meta-cluster". Specialized Grid Information and Control servers are deployed in order to translate from the abstract meta-interface into the specific interface used by each cluster. Grid interfaces and protocols have not been standardized. After an intense period of development during the 2000s, a handful of partially compatible Grid schemes are now deployed. Many claims of success in providing resources on the Grid

[68] "The grid: blueprint for a new computing infrastructure", Ian Foster and Carl Kesselman, editors. San Francisco: Morgan Kaufmann Publishers, c1999. ISBN 1558604758.

come from providing simple CPU-intensive services with little input or output. Providing reliable data-centric services has proven to be much more difficult, both from the point of view of the operation of resource centers themselves and of the Information and Control servers which provide the Grid interface for each center. Nevertheless, the CERN Computer Center, the WLCG Tier-1 centers and a subset of the WLCG Tier-2 centers do provide the needed reliable data-centric services needed for nucleating the rest of resources (see below).

The European Grid Infrastructure (EGI) program,[69] financed in part by the EU, has deployed the largest production Grid Infrastructure to date in collaboration with the Worldwide LHC Computing Grid (see next section). Other Grid Infrastructures provide support for particle and astroparticle physics, for example the Open Science Grid (OSG) in the USA. The EGI Grid will be described as an example of a working Grid used for daily work by particle physicists. EGI defines a Grid Infrastructure as a set of services deployed over Internet that allows a large number of Resource Centers, each with a different security and management domain, to be used in a coherent fashion by a large number of users from different institutions grouped in virtual organizations. Usually, a virtual organization corresponds to a project.

The basic services offered by the EGI Grid Infrastructure are based on the UMD middleware distribution,[70] which is deployed in a managed way, including monitoring and fault detection, with the support of a distributed operations organization provided by resource centers, many of which are associated to LHC computing or national Research and Education networks. Resource centers deploy Remote Data Servers using a variety of Disk Servers (possibly with magnetic tape back ends) which expose a uniform interface to the Grid via UMD, whose instances are called "Storage Elements". Authorization and access control is accomplished using the certificate method described earlier with tokens generated by the Virtual Organization Membership Service (VOMS) upon presentation of X.509 certificates issued by national research agencies and declaration of the virtual organization (e.g. project) under which the user wishes to be authorized. The actual data flow is predominantly through the *ftp* protocol, wrapped in tools such as *gridftp* (which incorporates certificate authentication and multiple parallel transmission streams).[71]

Each resource center implements a "Computing Element" (CE) service, a set of Information and Control Servers which map a specific batch system to an abstract

[69] See http://www.egi.eu. EGI operates the international coordination structure for Grid resource centers across Europe. EGI also coordinates the maintenance and security support for the Grid software originally developed in the 2000s by the EU Enabling Grids for E-sciencE (EGEE) series of projects and the EU DataGrid project.

[70] The Unified Middleware Distribution (UMD) maintained by EGI is based on software developed in the EU European Middleware Infrastructure (EMI) project. The EMI software is the result of integration and compatibility work done on the gLite middleware, developed within the EGEE Project, the ARC middleware, developed by NorduGrid, and various other packages. See https://wiki.egi.eu/wiki/Middleware

[71] A more sophisticated Grid data access tool known as *srm* was used in the past but is largely being abandoned due to maintenance and management issues.

Grid batch system. Directives may be given to the CE specifying job requirements such as minimum amount of memory, length of execution, etc. A challenge in configuring very large Grids is to reach agreement on the normalizations used to measure cluster resources, for example the speed of processors. Users submit their jobs through an additional server, called a Resource Broker, which accesses information in tables maintained by all available CEs and submits the jobs as appropriate to balance the load amongst the multiple sites conforming the Grid. Large projects, such as the LHC experiments, have developed job management services which in practice replace the Resource Brokers, interacting directly with CEs.

The Grid data access services described above offer functionality similar to the low levels of operating systems, but with the great advantage of generating coherent behavior across many Resource Centers. In practice, these services are often too low-level for practical use by applications. Hence, additional software, services and APIs are provided by middleware distributions at a higher level. One example is the File Transfer Service (FTS),[72] designed to support applications which require automated management, monitoring and error correction of replication of large numbers of files and large volumes of data, such as the LHC experiments. It uses a database back end to store replication requests and instances of finite state machines to reliably accomplish the corresponding data replications, including error correction and reporting. In addition, FTS implements the concept of a channel (a virtual path between two Resource Centers) which may be managed by operators, adjusting parameters such as the number of simultaneous files being transferred, the maximum bandwidth used, etc.

Large scientific projects, such as particle physics experiments, can have specific needs that go beyond the scope or functionality of basic middleware (such as UMD) and basic Grid services (such as those offered by EGI or OSG). In these cases, the projects themselves provide their own enhancements.[73]

14.7 The Worldwide LHC Computing Grid (WLCG)

The largest data processing capacity deployed so far using clusters linked via Grid infrastructures is the Worldwide LHC Computing Grid (WLCG or LCG).[74] WLCG links some 170 clusters located in 42 different countries into a single

[72] Andrey Kiryanov, Alejandro Alvarez Ayllon and Oliver Keeble, "FTS3/WebFTS – A Powerful File Transfer Service for Scientific Communities", Procedia Computer Science, Volume 66, 2015, Pages 670–678, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.11.076

[73] Readers should inquire within their own projects about these enhancements. They usually fall into two categories: Job management enhancements, for example the DIRAC package used by the LHCb experiment, and File or Dataset management enhancements, such as the AAA package used in CMS and the Rucio package used by ATLAS.

[74] See http://www.cern.ch/lcg

Table 14.1 Capacity deployed in the tiers of the LHC computing grid in 2018

		2018
CPU (KHS06[a])	Tier-0[b]	1272
	All Tier-1	2061
	All Tier-2	2562
Disk (TeraBytes)	Tier-0	90700
	All Tier-1	192214
	All Tier-2	185455
Tape (TeraBytes)	Tier-0	284700
	All Tier-1	461050

[a]Thousands of HEPSPEC06 units. HEPSPEC06 is a measure of CPU speed maintained by the particle physics community. See http://w3.hepix.org/benchmarking.html. As a very rough guideline, a current 2 GHz CPU delivers about 15 HS06

[b]In addition to the Tier-0, CERN also deploys an Analysis Facility (CAF) which is not included in this table

coherent distributed system which supports all data processing operations of the LHC experiments. WLCG defines a global service supported in part by Grid Infrastructure services in various countries and regions, for example EGI in Europe and OSG in the USA. The large amount of data handled by WLCG (over 100 Petabytes per year exchanged between sites) requires the sites to be organized in three "Tiers". Centers belonging to a given Tier are dedicated to perform certain specific tasks within the computing model of a given LHC experiment. There is one Tier-0 center at CERN, which handles Data Recording, Prompt Reconstruction and Remote Data Distribution (mostly to the Tier-1 centers). There are 13 Tier-1 centers, which perform Data Recording (between them they hold a second copy of all raw data stored at the Tier-0), Reprocessing, Data Reduction and Event Filtering. Hundreds of Tier-2 centers receive from the Tier-1 centers DST-type files (reduced and filtered) which physicists then use for analysis. Tier-2 centers also run event simulation, transmitting the outputs to Tier-1 centers for data recording and reconstruction. FTS, coupled to some experiment specific tools, is used to automate all data transmission.[75] The capacities deployed are quite large: about 400 thousand computing cores, 250 Petabytes of disk and nearly an Exabyte of tape. Table 14.1 shows the capacities deployed in 2018. These capacities have grown at an impressive rate of about 37% year-to-year since 2008. An even larger growth is expected for Run-3 of the LHC starting in 2021.

[75]The Wide Area Network requirements between the Tier-0 center and the 13 Tier-1 centers are so demanding that a special network based on point-to-point 10 Gbps links has been deployed (the LHC Optical Private Network, or LHCOPN). As of this writing, the LHCOPN links are being upgraded to 100 Gbps.

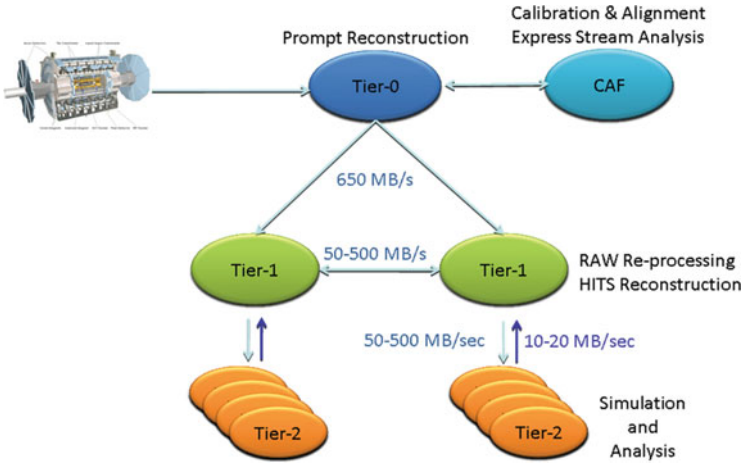


Fig. 14.2 Schematic view of the original ATLAS computing model

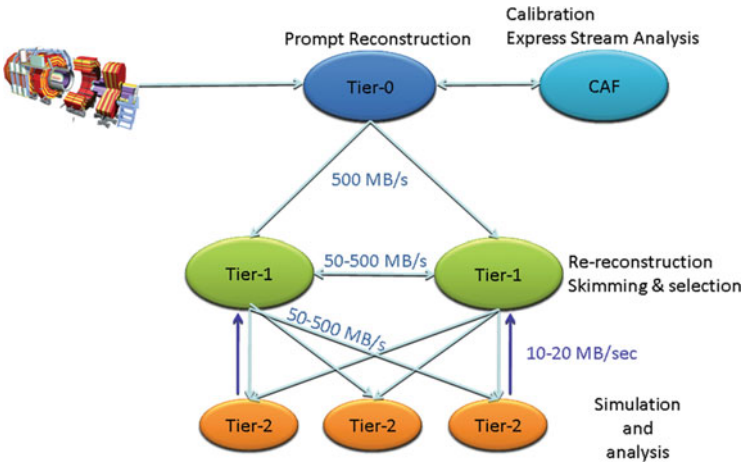


Fig. 14.3 Schematic view of the original CMS computing model

As mentioned earlier, the detailed breakdown of tasks performed at each Tier depends on the computing model of each experiment. On the other hand, the Tier-0 and most Tier-1 sites serve multiple experiments, whereas most Tier-2 sites only serve a single experiment. This leads to quite complex optimization issues, especially in regards to data transmission, which depending on the experiment may involve few or many site pairings. For example, Figs. 14.2 and 14.3 show schematically the original computing models for the ATLAS and CMS experiments, and the different connectivity patterns between Tiers can clearly be seen. Originally, ATLAS closely linked a given Tier-1 to a group of Tier-2 sites, and hence the data flows in a more "hierarchical" fashion. The original CMS model had a focus on

coherent operation of all Tier-1 sites and allows data to flow between any pairing of Tier-1 and Tier-2 centers. The actual situation is even more complex, as there can be large size differences between Tier-1 or Tier-2 centers at different sites.

The computing models of the LHC experiments have undergone substantial evolution in the last 10 years, influenced by the operational experience that has been gained. The most important lesson is that reliably managing storage at sites is a difficult task, requiring specialized skills not always available at resource centers. In addition, the management of data and replica placement by the central production teams of each experiment requires substantial human resources, and the effort needed scales with the number of sites as opposed to the overall amount of storage. Hence, a model where a limited number of large sites provide highly reliable 10–100 Petabyte storage platforms, with the rest of the sites providing CPU power with input–output via Remote Data Servers, is globally more manageable and economical.

The original, static Tier hierarchy was motivated in part by the high costs and technical difficulties of deploying Gigabit per second (Gbps) WAN networking. However, networking in general and WAN in particular have had the fastest evolution of all computer industry technologies. In addition, Research and Education Networks provide a wider variety of connection options, including 10 and 100 Gbps links for specially demanding sites. This means that currently some Tier-2 sites have equal networking capabilities to a Tier-1 site, whereas they may or may not achieve the reliability levels of a Tier-1 site. It is convenient, then, to use these sites in a more flexible manner through a more dynamic hierarchical model.

Hence, the WLCG model has evolved towards differentiating two types of sites: Nucleation sites, which have high reliability and provide Petabyte-level storage, and satellite sites, which are mostly expected to deliver CPU capacity. WAN based input–output is accomplished through the custom *xrootd* or the standard *http* protocols, and the storage is remotely managed by the central experiment project teams via standard *WebDAV* protocols interacting with the Remote Data Servers at nucleation sites. Sites are continuously monitored and their status as nucleation points is defined dynamically according to their current reliability. In this model, most of the Tier-1 sites and larger Tier-2 sites act as nucleation sites. A site with reliability or availability problems will be temporarily removed as a nucleation site and restored when monitoring reveals the issues have been resolved. Tier-1 sites continue to have the differentiating feature of operating the robotic magnetic tape libraries used for bulk-archiving of the data.

14.8 Current State-of-the-Art Technologies

The computing industry continues to evolve at a rapid pace. The decade of the 2010s has seen the consolidation of a number of technologies and the emergence of an important change in business models for the provision of computing services. These are being integrated into the computing environments for particle physics.

14.8.1 Multi-core CPUs and Virtualization

Multi-core CPUs have been discussed in Sect. 14.1. Ideally, all applications would by now be multi-threaded and capable of efficiently utilizing the full capacity of the cores in a CPU. Unfortunately, thread-safe programming is a specialized skill and the vast majority of programs used in the world are not thread-safe. Hence, a different approach is needed to fully use multi-core CPUs.

In addition, applications running in distributed computing environments often have intricate dependencies on the underlying libraries and on the operating system. Continuously porting applications to newer versions of the underlying environment and verifying their proper operation requires much larger human resources than can be afforded by scientific communities (and even by many business communities). Hence, deployment of applications in these environments enters into conflict with system administration good practices, which require deploying new operating system versions soon after they become available.

These two factors have lead to the re-emergence of two techniques from the past which enable operating system virtualization:[76] hypervisors and virtual machines (first deployed in the 1970s by IBM in its VM operating system[77]), and containers (also known as jails, partitions or zones, first deployed commercially in the early 2000s by Sun Microsystems[78]).

Virtual machines are implemented by deploying over the hardware a special type of operating system called a hypervisor. The hypervisor is designed to efficiently run multiple copies of a special type of application called a virtual image. In addition, newer versions of hardware implement special features to accelerate virtual image execution. The virtual image contains one or more applications together with the full operating system and library environment required by them. In this way, every time the hypervisor starts running an image it effectively bootstraps a new virtual computer, which is referred to as a virtual machine. Each image can have a different operating system and different libraries. If properly deployed, virtual machines allow to run older, less secure operating systems without compromising security, as the hypervisor can limit access to hardware and file system resources.

Virtual machines allow the implementation of many interesting features. One of the best known is live-migration,[79] which combines virtual machines with checkpoint/restart and high performance multi-connection file servers to be able

[76]For a short, rather technical overview, see "Virtualization and Containerization of Application Infrastructure: A Comparison", Thijs Scheepers, 21st Twente Student Conference on IT June 23rd, 2014, Enschede, The Netherlands. Available from the author at https://thijs.ai/papers/scheepers-virtualization-containerization.pdf

[77]"The Origin of the VM/370 Time-Sharing System", R. J. Creasy, IBM J. Res. Develop., Vol. 25, No. 5 (September 1981).

[78]Price, Daniel, and Andrew Tucker. "Solaris Zones: Operating System Support for Consolidating Commercial Workloads." LISA. Vol. 4. 2004.

[79]Sapuntzakis, Constantine P., et al. "Optimizing the migration of virtual computers." ACM SIGOPS Operating Systems Review 36.SI (2002): 377–390.

to move virtual machines between different hardware units. This can be useful for providing uninterruptable services or for saving energy in environments with large load variations.

Traditional offline batch computing in particle physics, however, predominantly uses virtualization for simply packaging together an application and its environment. This reveals a weakness of traditional virtualization, as the many copies of operating systems in the virtual machines require large amounts of memory (RAM) to execute efficiently, increasing the costs of the computing clusters. This has recently been circumvented by a new virtualization package, *Singularity*,[80] which has been rapidly adopted by the scientific community, and large portions of WLCG now support the execution of *Singularity* images. An additional advantage of *Singularity* is that it can be easily supported on supercomputers.

Containers, in their present form, represent a slightly different approach to virtualization. The current de-facto standard is *Docker*,[81] which is widely used in commercial applications. Containers are closely associated to Linux operating systems and to Cloud Computing environments (see below). Containers implement virtual machines which are less hermetic than traditional ones and which have certain limitations. On the other hand, containers can be rapidly created and destroyed, and packages exist, such as *Kubernetes*,[82] for orchestrating a set of containers which together can implement sophisticated applications using micro-service architectures.

In a somewhat unexpected way, *Singularity* and container environments have recently emerged as the leading approach for preserving data processing environments in order to ensure reproducibility of scientific results.

14.8.2 Cloud Computing and the Use of Commercial Data Processing Services

Cloud Computing[83] is the currently used term[84] to describe the deployment of computing resources and higher level computing services in a shareable, user configurable way. It is the new paradigm that replaced client-server computing in the

[80] Kurtzer GM, Sochat V, and Bauer MW (2017) Singularity: Scientific containers for mobility of compute. PLoS ONE 12(5): e0177459. https://doi.org/10.1371/journal.pone.0177459

[81] Boettiger, Carl. "An introduction to Docker for reproducible research." ACM SIGOPS Operating Systems Review 49.1 (2015): 71–79.

[82] Bernstein, David. "Containers and cloud: From lxc to docker to kubernetes." IEEE Cloud Computing 3 (2014): 81–84.

[83] Mell, Peter, Grance, Tim, "The NIST Definition of Cloud Computing", Special Publication 800-145, National Institute of Standards and Technology, U.S. Department of Commerce (September 2011) https://doi.org/10.6028/NIST.SP.800-145

[84] Regalado, Antonio. "Who coined 'cloud computing'." Technology Review 31 (2011). https://www.technologyreview.com/s/425970/who-coined-cloud-computing/Retrieved 3 February 2019.

2000s, starting a new cycle in the development of outsourced computing services. The first large scale commercialization was done in 2006, when Amazon launched the Elastic Compute Cloud (EC2) service. The first large-scale production cloud services in the academic domain were NASA's OpenNebula[85] and the services deployed in the RESERVOIR EU-funded project.[86]

Cloud computing is still under heavy evolution, and a full description of the technology is beyond the scope of this work. It deploys interfaces and toolkits that allow users to configure virtual computers, storage servers, clusters and even networks. The ultimate goal is to use software to define the characteristics of a distributed computing service and then automatically map it to the needed hardware. This process is called provisioning.

Cloud computing can be used to implement private research data centers. CERN has migrated essentially all of its data center platforms to be managed by the *OpenStack*[87] cloud management system. However, the benefits of introducing cloud-style management in smaller data centers are not evident at present, as standard clusters are probably sufficient to fulfill the needs. In addition, there is at present a lack of personnel trained in cloud computing deployment and operation.

Cloud computing technologies started in the academic world, but are now driven by the cloud computing industry, which has grown in a decade to have sales of over 200 G$/year. Cloud infrastructure companies deploy data centers which are much larger than academic computing centers, with the exception of a few academic supercomputing centers. Computer room floor areas above one hundred thousand square meters, power feeds above 50 MW[88] and capabilities for hosting tens of thousand of servers and close to a million cores are typical.

These huge deployments bring with them large economies of scale, making them attractive for scientific computing. In some cases, U.S. funding agencies have started to give grants for the purchase of commercial cloud computing services, replacing funding for purchasing research computing clusters. Gateways have been developed to include commercial clouds as part of WLCG, especially for CPU intensive tasks such as simulations. In the European Union, the HNSciCloud[89] pre-commercial procurement project in 2016–2018 was aimed to stimulate industry developments to enable hybrid clouds between research and commercial data centers capable of

[85]Nebula Cloud Computing Platform (20 November 2012) https://www.nasa.gov/open/nebula.html

[86]Rochwerger, Benny, et al. "The reservoir model and architecture for open federated cloud computing." IBM Journal of Research and Development 53.4 (2009): 4–1.

[87]https://www.openstack.org/software/

[88]For comparison, CERN's Meyrin data center power was 3.5 MW in 2018. "Data Centre: Key Info & Numbers". http://information-technology.web.cern.ch/sites/information-technology.web.cern.ch/files/CERNDataCentre_KeyInformation_November2018V1.docx.pdf. Retrieved 3 February 2019.

[89]Helix Nebula – The Science Cloud with Grant Agreement 687614 is a Pre-Commercial Procurement Action funded by H2020 Framework Programme. More information at https://hnscicloud.eu

executing data-intensive tasks. A similar project named ARCHIVER[90] will explore in 2019–2021 the commercial provision of cloud mass storage services for scientific data archiving and preservation.

Cloud computing deployment, especially on commercial services, is quite complex at the infrastructure and platform level. Fortunately, past investments in Grid computing can be re-used in order to hide these complexities from the vast majority of users. Job and data management packaged used for Grid computing have already been deployed on private and commercial clouds, as well as portals for analysis using higher level tools such as *Jupyter* notebooks and *Hadoop* data servers.

14.9 Future Challenges and Directions

Particle physics has historically been at the forefront of innovation in the adaptation of computing systems to fulfill its growing needs, as well as to accommodate tight budgets. It is rare, however, that the configurations developed in the particle physics context can directly be taken over by commercial systems. They should be considered as "precursors" or "early adoptions" of advances to come into general use. This situation has become a challenge with the growing time spans for experiments. If not properly managed, it may result in huge (often hidden) costs for maintenance and the lack of adoption of other innovations.

A first consideration for the future development is the number of scientific projects requiring wide-area distributed processing. In the short term, LHC will continue to produce unprecedented amounts of data and will be joined by a few experiments with similar needs, such as the Square Kilometer Array (SKA)[91] radiotelescope. However, many other experiments, such as neutrino detectors and astroparticle experiments, may produce orders of magnitude smaller data volumes, due to their own instrumental nature or to advances in data reduction within the data acquisition platforms. For example, even though 100 Gbps WAN will become common place by the 2020s, the baseline design of the Cherenkov Telescope Array (CTA)[92] requires only a 1 Gbps network interface for the sites (a capacity already available to home users in many countries). This is because CTA plans to use powerful clusters housed in compact containerized data centers placed at the sites to reduce the data at the instrument, thus avoiding the recording of large amounts of

[90]ARCHIVER – Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516. More information at https://archiver-project.eu

[91]Dewdney, Peter E., et al. "The square kilometre array." Proceedings of the IEEE 97.8 (2009): 1482–1496.

[92]Actis, M., et al. "Design concepts for the Cherenkov Telescope Array CTA: an advanced facility for ground-based high-energy gamma-ray astronomy." Experimental Astronomy 32.3 (2011): 193–316.

raw data. Developments in the LHC domain, such as the LHCb turbo stream,[93] may also reduce the need for large-scale wide-area distributed processing. In parallel, as discussed below, data centers, particularly commercial ones, are growing in capacity. It is quite conceivable that the future needs of many experiments could be serviced by just a few data centers.

Another issue is the foreseen disappearance of magnetic tape technology. Tape has already disappeared from the commodity market, driven by ever lower disk prices and alternative backup media such as DVDs. If electrical energy considerations are excluded, the overall cost of high-end tape, associated robotic mechanisms and periodic migration to new media is becoming less attractive. Nevertheless, as of this writing, tape storage continues to be more economical than disk storage, especially if electrical costs are taken into account. The prospects are worrisome, however, as fewer companies continue to do R&D in tape technology. Hence, the particle physics community should be preparing for a hypothetical tape-less future by adapting its data management and cluster architectures as described above, allowing for multiple disk copies of the same dataset on a cluster or across a Grid or in several Clouds.

Grid and Cloud computing have brought along a resurgence of "timesharing",[94] with research data centers deploying large clusters which give services to many projects of which only a minority are particle or astroparticle physics projects. The early adoption of Grids by particle physics, however, has resulted in development and deployment of a number of cluster and Grid tools which require extremely specific, non-commercially supported operating system or data server configurations which are of no interest to other projects or to the managers of general purpose clusters or Clouds. This does not mean that the developments led by particle physics are not worthwhile or of good quality; it simply means that further steps must be taken outside the particle physics scope to ensure a sustainable future and general applicability. Therefore, the particle physics community has to share knowledge about its developments, but also continuously evaluate alternative solutions. A case in point may be the use of non-standard data serving packages by LHC sites: it is perceived as the best solution (especially for LHC Tier-1 centers), but smaller particle or astroparticle projects with much lower data volumes often prefer what they perceive as a simpler, more standard solution such as NFS. Standard deployment of NFS on a cluster with thousands of CPU cores will however not be favored by system managers or commercial providers for a number of practical operational reasons, and they will propose newer technologies such as Ceph[95]

[93] Benson, Sean, et al. "The LHCb turbo stream." Journal of Physics: Conference Series. Vol. 664. No. 8. IOP Publishing, 2015.

[94] The term "timesharing" was coined in the 1960s to describe the simultaneous use of a single "mainframe" computer for multiple batch or interactive tasks by multiplexing processes into the mainframe CPU.

[95] Weil, Sage A., et al. "Ceph: A scalable, high-performance distributed file system." Proceedings of the 7th symposium on Operating systems design and implementation. USENIX Association, 2006.

or OpenStack swift[96] Object Storage. Therefore, a likely scenario is that current storage access techniques used by particle physics will become unsupported in the future. Of course, virtualization and Cloud techniques will make it possible for the particle physics community to deploy these legacy technologies themselves, with the corresponding expenses in personnel. A better way forward is to collaborate with industry, and between various Resource Centers, to test alternative tools and methods. Realistic evaluations are far from trivial, however, as they may require substantial investments by all parties, for example to provide UMD Grid or Cloud interfaces.

Another looming issue is the efficiency of use of the CPUs under data-intensive conditions. One basic assumption of the distributed architecture used by the particle physics community, as described above, is the existence of an infinitely powerful network connecting all elements, which in effect assumes infinitely powerful data servers as well as data ingestion by CPUs. Ever increasing data processing requirements and more and more powerful multi-core CPU nodes, however, are revealing data serving performance bottlenecks in clusters coming from the network and disk and CPU server limitations. Removing these bottlenecks may require more sophisticated architectures to be deployed.

A much more severe issue is related to the simplistic, often naïve, manner in which large particle physics projects pretend to use Grids and Clouds. They essentially desire them to behave as a very large cluster perfectly tuned to their needs, ignoring that imposing unreasonably high requirements in peak network and data serving rates within and between clusters can be very expensive. A Grid would better be viewed as a loosely connected federation of largely autonomous, self-sufficient clusters, with some of these clusters possibly hosted on commercial services.

These and other future challenges are being addressed in a coherent manner by the particle physics community through the HEP Software Foundation[97,98] and will certainly require a vigorous new cycle of research and development, in collaboration with computer scientists and engineers, the supercomputing and HPC communities and industry.

[96] Arnold, Joe. Openstack swift: Using, administering, and developing for swift object storage. "O'Reilly Media, Inc.", 2014.

[97] Alves Jr, Antonio Augusto. A Roadmap for HEP Software and Computing R&D for the 2020s. No. HSF-CWP-2017-001; HSF-CWP-2017-01; FERMILAB-PUB-17-607-CD; arXiv: 1712.06982. Fermi National Accelerator Lab. (FNAL), Batavia, IL (United States); Brookhaven National Laboratory (BNL), Upton, NY (United States); Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States); SLAC National Accelerator Lab., Menlo Park, CA (United States); Thomas Jefferson National Accelerator Facility (TJNAF), Newport News, VA (United States); Argonne National Lab. (ANL), Argonne, IL (United States), 2017.

[98] Slides from the Computing in High Energy Physics 2018 conference (proceedings to be published) https://indico.cern.ch/event/587955/contributions/3012294/attachments/1681524/2708636/CHEP18_-_CWP_Lessons_and_Future_Work.pdf

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 15
Statistical Issues in Particle Physics

Louis Lyons

15.1 Introduction

In recent years there has been a growing awareness by particle physicists of the desirability of using good statistical practice. This is because the accelerator and detector facilities have become so complex and expensive, and involve so much physicist effort to build, test and run, that it is clearly important to treat the data with respect, and to extract the maximum information from them. The PHYSTAT series of Workshops and Conferences[1–12] has been devoted specifically to statistical issues in particle physics and neighbouring fields, and many interesting articles can be found in the relevant Proceedings. These meetings have benefited enormously from the involvement of professional statisticians, who have been able to provide specific advice as well as pointing us to some techniques which had not yet filtered down to Particle Physics analyses.

Analyses of experimental data in Particle Physics have, perhaps not surprisingly, tended to use statistical methods that have been described by other Particle Physicists. There are thus several books written on the subject by Particle Physicists[13]. The Review of Particle Physics properties[14] contains a condensed review of Statistics.

Another source of useful information is provided by the statistics committees set up by some of the large collaborations (see, for example, refs. [15–18]). Some conferences now include plenary talks specifically on relevant statistical issues (for example, Neutrino 2017[19], NuPhys17 and NuPhys18[20]), and the CERN

L. Lyons (✉)
Physics, University of Oxford, Oxford, UK
e-mail: louis.lyons@physics.ox.ac.uk

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_15

Summer Schools for graduate students regularly have a series of lectures on statistics for Particle Physics[21].

This article is a slightly updated version of the one that appeared in ref. [22] in 2012.

15.1.1 Types of Statistical Analysis

There are several different types of statistical procedures employed by Particle Physicists:

- Separating signal from background: Almost every Particle Physics analysis uses some method to enhance the possible signal with respect to uninteresting background.
- Parameter determination: Many analyses make use of some theoretical or empirical model, and use the data to determine values of parameters, and their uncertainties and possible correlations.
- Goodness of fit: Here the data are compared with a particular hypothesis, often involving free parameters, to check their degree of consistency.
- Comparing hypotheses: The data are used to see which of two hypotheses is favoured. These could be the Standard Model (SM), and some specific version of new physics such as the existence of SUperSYmmetry (SUSY), or the discovery of the Higgs boson[23].
- Decision making: Based on one's belief about the current state of physics, the value of possible discoveries and estimates of the difficulty of future experiments, a decision is made on what should be thrust of future research. This subject is beyond the scope of this article.

15.1.2 Statistical and Systematic Uncertainties

In general any attempt to measure a physics parameter will be affected by statistical and by systematic uncertainties. The former are such that, if the experiment were to be repeated, random effects would result in a distribution of results being obtained. These can include effects due to the limited accuracy of the measurement devices and/or the experimentalist; and also from the inherent Poisson variability of observing a number of counts n. On the other hand, there can be effects that shift the measurements from their true values, and which need to be corrected for; uncertainties in these corrections contribute to the systematics. Another systematic effect could arise from uncertainties in theoretical models which are used to interpret the data. Scientists' systematics are often 'nuisance parameters' for statisticians.

Consider an experiment designed to measure the temperature at the centre of the sun by measuring the flux of solar neutrinos on earth. The main statistical

uncertainty might well be that due to the limited number of neutrino interactions observed in the detector. On the other hand, there are likely to be systematics from limited knowledge of neutrino cross-sections in the detector material, the energy calibration of the detector, neutrino oscillation parameters, models of energy convection in the sun, etc. If some calibration measurement or subsidiary experiment can be performed, this effectively converts a systematic uncertainty into a statistical one. Whether this source of uncertainty is quoted as statistical or systematic is not crucial; what is important is that possible sources of correlation between uncertainties here and in other measurements (in this or in other experiments) are well understood.

The magnitude of systematic effects in a parameter-determination situation can be assessed by fitting the data with different values of the nuisance parameter(s), and seeing how much the result changes[1] when the nuisance parameter value is varied by its uncertainty. Alternatively the nuisance parameter(s) for systematic effects can be incorporated into the likelihood or χ^2 for the fit; or a Bayesian method involving the prior probability distribution for the nuisance parameter can be used. (See Sects. 15.4.5 and 15.7.6 for ways of incorporating nuisance parameters in upper limit and in p-value calculations respectively).

How to assess systematics was much discussed at the first Banff meeting[6] and at PHYSTAT-LHC[24–26]. A special session of the recent PHYSTATν meeting at CERN[12] was devoted to systematics. Many reviews of this complex subject exist and can be traced back via ref. [27].

In general, much more effort is involved in estimating systematic uncertainties than for parameter determination and the corresponding statistical uncertainties; this is especially the case when the systematics dominate the statistical uncertainty.

Cowan[35] has considered the effect of having an uncertainty in magnitude of a systematic effect. As Cox has remarked[36], there is a difference in knowing that a correction has almost precisely a 20% uncertainty, or that it is somewhere between 0% and 40%.

15.1.3 Bayes and Frequentism

These are two fundamental approaches to making inferences about parameters or whether data support particular hypotheses. There are also other methods which do not correspond to either of these philosophies; the use of χ^2 or the likelihood are examples.

Particle physicists tend to favour a frequentist method. This is because in many cases we really believe that our data are representative as samples drawn according to the model we are using (decay time distributions often are exponential; the counts

[1] If the simulation yields a change in the result of $a \pm b$, there is much discussion about how the contribution to the systematic uncertainty should be assessed in terms of a and b—see ref. [27].

in repeated time intervals do follow a Poisson distribution; etc.), and hence we want to use a statistical approach that allows the data "to speak for themselves", rather than our analysis being sensitive to our assumptions and beliefs, as embodied in the assumed Bayesian priors. Bayesians would counter this by remarking that frequentist inference can depend on the reference ensemble, the ordering rule, the stopping rule, etc.

With enough data, the results of Bayesian and frequentist approaches usually tend to agree. However, in smallish data samples numerical results from the two approaches can differ.

15.1.3.1 Probability

There are at least three different approaches to the question of what probability is. The first is the mathematical one, which is based on axioms e.g. it must lie in the range 0–1; the probabilities of an event occurring and of it not occurring add up to 1; etc. It does not give much feeling for what probability is, but it does provide the underpinning for the next two methods.

Frequentists, not surprisingly, define probability in terms of frequencies in a long series of essentially identical repetitions[2] of the relevant procedure. Thus the probability of the number 5 being uppermost in throws of a die is 1/6, because that is the fraction of times we expect (or approximately observe) it to happen. This implies that probability cannot be defined for a specific occurrence (Will the first astronaut who lands on Mars return to earth alive?) or for the value of a physical constant (Does Dark Matter contribute more than 25% of the critical density of the Universe?).

In contrast, Bayesians define probability in terms of degree of belief. Thus it can be used for unique events or for the values of physical constants. It can also vary from person to person, because my information may differ from yours. The numerical value of the probability to be assigned to a particular statement is determined by the concept of a 'fair bet'; if I think the probability (or 'Bayesian credibility') of the statement being true is 20%, then I must offer odds of 4-to-1, and allow you to bet in either direction.

This difference in approach to probability affects the way Bayesians and frequentists deal with statistical procedures. This is illustrated below by considering parameter determination.

15.1.3.2 Bayesian Approach

The Bayesian approach makes use of Bayes' Theorem:

$$p(A|B) = p(B|A) \times p(A)/p(B), \tag{15.1}$$

[2]Bayesians attack this concept of 'essentially identical trials', claiming that it is hard to define it without using the concept of probability, thus making the definition circular.

where $p(A)$ is the probability or probability density of A, and $p(A|B)$ is the conditional probability for A, given that B has happened. This formula is acceptable to frequentists, provided the probablities are frequentist probabilities. However Bayesians use it with $A =$ parameter (or hypothesis) and $B =$ data. Then

$$p(parameter|data) \propto p(data|parameter) \times p(parameter), \tag{15.2}$$

where the three terms are respectively the Bayesian posterior, the likelihood function and the Bayesian prior. Thus Bayes' theorem enables us to use the data (as encapsulated in the likelihood) to update our prior knowledge ($p(parameter)$); the combined information is given by the posterior.

Frequentists object to the use of probability for physical parameters. Furthermore, even Bayesians agree that it is often hard to specify a sensible prior. For a parameter which has been well determined in the past, a prior might be a gamma function or log-normal or a (possibly truncated) Gaussian distribution of appropriate central value and width, but for the case where no useful information is available the choice is not so clear; it is easier to parametrise prior knowledge than to quantify prior ignorance. The 'obvious' choice of a uniform distribution has the problem of being not unique (Should our lack of knowledge concerning, for example, the mass of a neutrino m_ν be parametrised by a uniform prior for m_ν or for m_ν^2 or for $\log m_\nu$, etc?). Also a uniform prior over an infinite parameter range cannot be normalised. For situations involving several parameters, the choice of prior becomes even more problematic.

It is important to check that conclusions about possible parameter ranges are not dominated by the choice of prior. This can be achieved by changing to other 'reasonable' priors (sensitivity analysis); or by looking at the posterior when the data has been removed.

15.1.3.3 Frequentist Approach: Neyman Construction

The frequentist way of constructing intervals completely eliminates the need for a prior, and avoids considering probability distributions for parameters. Consider a measurement x which provides information concerning a parameter μ. For example, we could use a month's data from a large solar neutrino detector (x) to estimate the temperature at the centre of the sun (μ). It is assumed that enough is known about solar physics, fusion reactions, neutrino properties, the behaviour of the detector, etc. that, for any given value of μ, the probability density for every x is calculable. Then for that μ, we can select a region in x which contains, say, 90% of this probability. If we do this for every μ, we obtain a 90% confidence band; it shows

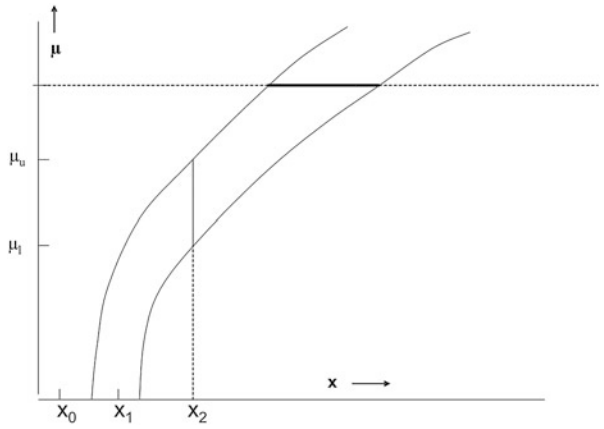


Fig. 15.1 The Neyman construction for setting a confidence range on a parameter μ. At any value of μ, it is assumed that we know the probability density for obtaining a measured value x. (For example, μ could be the temperature of the fusion reactor at the centre of the Sun, while α is the solar neutrio flux, estimated by operating a large underground solar neutrino detector for 1 month.) We can then choose a region in x which contains, say, 90% of the probability; this is denoted by the solid part of the horizontal line. By repeating this procedure for all possible μ, the band between the curved lines is constructed. This confidence band contains the likely values of x for any μ. For a particular measured value x_2, the confidence interval from μ_l to μ_u gives the range of parameter values for which that measured value was likely. For x_2, this interval would be two-sided, while for a lower value x_1, an upper limit would be obtained. In contrast, there are no parameter values for which x_0 is likely, and for that measured value the confidence interval would be empty

the values of x which are likely results[3] of the experiment for any μ, assuming the theory is correct (see Fig. 15.1). Then if the actual experiment gives a measurement x_2, it is merely necessary to find the values of μ for which x_2 is in the confidence band. This is the Neyman construction.

Of course, the choice of a region in x to contain 90% of the probability is not unique. The one shown in Fig. 15.1 is a central one, with 5% of the probability on either side of the selected region. Another possibility would be to have a region with 10% of the probability to the left, and then the region in x extends up to infinity. This choice would be appropriate if we always wanted to quote upper limits on μ. Other choices of 'ordering rule' are also possible (see, for example, Sect. 15.4.3).

The Neyman construction can be extended to more parameters and measurements, but in practice it is very hard to use it when more than two or three parameters are involved; software to perform a Neyman construction efficiently in several dimensions would be very welcome. The choice of ordering rule is also very important. Thus from a pragmatic point of view, even ardent frequentists

[3]The adjective 'likely' is appropriate for central intervals. For upper limits on μ, however, the accepted values of x for a given μ extend to infinity, and so 'preferred results for the given ordering rule' would be more appropriate.

are prepared to use Bayesian techniques for multidimensional problems (e.g. with systematics). They would, however, like to ensure that the technique they use provides parameter intervals with reasonable frequentist coverage.

15.1.3.4 Coverage

One of the major advantages of the frequentist Neyman construction is that it guarantees coverage. This is a property of a statistical technique[4] for calculating intervals, and specifies how often the interval contains the true value μ_t of the parameter. This can vary with μ_t.

For example, for a Poisson counting experiment with parameter μ and observed number n, a (not very good) method for providing an interval for μ is $n \pm \sqrt{n}$. Thus an observed $n = 2$ would give a range 0.59–3.41 for μ. If $\mu = 2.01$, observed values $n = 2,\ 3$ and 4 result in intervals that include $\mu = 2.01$, while other values of n do not. The coverage of this procedure for $\mu = 2.01$ is thus the sum of the Poisson probabilities for having $n = 2,\ 3$ or 4 for the given μ.

For a discrete observable (e.g. the number of detected events in a search for Dark Matter), there are jumps in the coverage; in order to avoid under-coverage, there is necessarily some over-coverage. However, for a continuous observable (e.g. the estimated mass of the Higgs boson) the coverage can be exact.

Coverage is not guaranteed for methods that do not use the Neyman construction (see Sect. 15.2.1). Interesting plots of coverage as a function of the parameter value for the simple case of a Poisson counting experiment can be found in ref. [32].

15.1.3.5 Likelihoods

The likelihood approach makes use of the probability density function (pdf) for observing the data, evaluated for the data actually observed.[5] It is a function of any parameters, although it does not behave like a probability density for them. It provides a method for determining values of parameters. These include point estimates for the 'best' values, and ranges (or contours in multi-parameter situations) to characterise the uncertainties. It usually has good properties asymptotically, but a major use is with sparse multi-dimensional data.

The likelihood method is neither frequentist nor Bayesian. It thus does not guarantee frequentist coverage or Bayesian credibility. It does, however, play a central role in the Bayesian approach, which obtains the posterior probability

[4]It is important to realise that coverage is a property of the **method**, and not of an **individual measurement**.

[5]The pdf $f(x, \mu_0)$ gives the probability density for obtaining various data x when the parameter has some specified value μ_0. The likelihood is the same function of two variables $f(x_0, \mu)$, but now with x_0 fixed at the data actually obtained, and μ regarded as the variable.

density by multiplying the likelihood by the prior. The Bayesian approach thus obeys the likelihood principle, which states that the only way the experimental data affects inference is via the likelihood function. In contrast, the Neyman construction requires not only the likelihood for the actual data, but also for all possible data that might have been observed.

Because the likelihood is not a probability density, it does not transform like one. Thus the value of the likelihood for a parameter μ_0 is identical to that for $\lambda_0 = 1/\mu_0$. This means that ratios of likelihoods (or differences in their logarithms) are useful to consider, but that the integration of tails of likelihoods is not a recognised statistical procedure.

A longer account of the Bayesian and frequentist approaches can be found in ref. [28]. Reference [29] provides a very readable account for a Poisson counting experiment.

15.2 Likelihood Issues

In this section, we discuss some potential misunderstandings of likelihoods.

15.2.1 $\Delta(lnL) = 0.5$ *Rule*

In the maximum likelihood approach to parameter determination, the best value λ_0 of a parameter is determined by finding where the likelihood maximises; and its uncertainty is estimated by finding how much the parameter must be changed[6] in order for the logarithm of the likelihood to decrease by 0.5 as compared with the maximum.[7] From a frequentist viewpoint, this should ideally result in the parameter range having 68% coverage. That is, in repeated use of this procedure to estimate the parameter, 68% of the intervals should contain the true value of the parameter, whatever its true value happens to be.

If the measurement is distributed about the true value as a Gaussian with constant width, the likelihood approach will yield exact coverage, but in general this is not so. For example, Garwood[31] and Heinrich[32] have investigated the properties of the likelihood approach (and other methods too) to estimate μ, the mean of a Poisson, when n_{obs} events are observed. Because n_{obs} is a discrete variable, the coverage is

[6] If there are more than just one parameter, the likelihood must of course be remaximised with respect to all the other parameters when looking for the $\Delta(lnL) = 0.5$ points. Alternatively, a region in multi-parameter space can be selected by finding the contour at which $\Delta(lnL)$ decreases from its maximum by an amount which depends on the number of parameters.

[7] This (like several other methods) can give rise to asymmetric uncertainties. Techniques for dealing with this have been discussed by Barlow[30].

a discontinuous function of μ, and varies from 100% at $\mu = 0$ down to 30% at $\mu \approx 0.5$.[8]

15.2.2 Unbinned Maximum Likelihood and Goodness of Fit

With sparse data, the unbinned likelihood method is a good one for estimating parameters of a model. In order to understand whether these estimates of the parameters are meaningful, we need to know whether the model provides an adequate description of the data. Unfortunately, as emphasised by Heinrich[33], the magnitude of the unbinned maximum likelihood is often independent of whether or not the data agree with the model. He illustrates this by the example of the determination of the lifetime τ of a particle whose decay distribution is $(1/\tau)\exp(-t/\tau)$. For a set of observed times t_i, the maximum likelihood L_{max} depends on the data t_i only through their average value $\bar{t}$. Thus any data distributions with the same $\bar{t}$ would give identical L_{max}, which demonstrates that, at least in this case, L_{max} gives no discrimination about whether the data are consistent with the expected distribution.

Another example is fitting an expected distribution $(1 + \alpha\cos^2\theta)/(1 + \alpha/3)$ to data θ_i on the decay angle of some particle, to determine α. According to the expected functional form, the data should be symmetrically distributed about $\cos\theta = 0$. However, the likelihood depends only on the **square** of $\cos\theta$, and so would be insensitive to all the data having $\cos\theta_i$ negative; this would be very inconsistent with the expected symmetric distribution.

In contrast Baker and Cousins[34] provide a likelihood method of measuring goodness of fit for a data **histogram** compared to a theory. The Poisson likelihood $P_{Pois}(n|\mu)$ for each bin is compared with that for the best possible predicted value $\mu_{best} = n$ for that bin. Thus the Baker-Cousins likelihood ratio

$$LR_{BC} = \Pi \frac{e^{-\mu_i}\mu_i^{n_i}/n_i!}{e^{-n_i}n_i^{n_i}/n_i!} = \Pi e^{(n_i-\mu_i)}(\mu_i/n_i)^{n_i} \tag{15.3}$$

is such that asymptotically $-2lnLR_{BC}$ is distributed as χ^2.[9] For small μ, the Baker-Cousins likelihood ratio is better than a weighted sum of squares for assessing goodness of fit.

[8] It is of course not surprising that methods that are expected to have good asymptotic behaviour may not display optimal properties for $\mu \approx 0$.

[9] The binned Poisson likelihood is not a measure of fit. This is because, for example. $\mu_i = n_i = 1$ and $\mu_i = n_i = 100$ both correspond to perfect agreement between data and prediction, but $P_{Pois}(1|1.0)$ is much larger than $P_{Pois}(100|100.0)$.

15.2.3 Profile Likelihood

In many situations the likelihood is a function not only of the parameter of interest ϕ but also other parameters. These may be other physics parameters (for example, in neutrino oscillation experiments where the mixing angles and differences in mass-squared of the various neutrinos are relevant), but can also be nuisance parameters ν associated with systematic effects (e.g. jet energy scales, particle identification efficiencies, etc.). To make statements about ϕ, the likelihood $L(\phi.\nu)$ is often 'profiled' over the nuisance parameters, i.e. at each value of ϕ, the likelihood is remaximised with respect to ν. Thus

$$L_{prof}(\phi) = L(\phi, \nu_{max}(\phi)) \tag{15.4}$$

Then $L_{prof}(\phi)$ is used much as the ordinary likelihood when there are no nuisance parameters.

A profile likelihood is in general wider than the likelihood for a fixed value of the nuisance parameter ν; this results in the uncertainty in the parameter of interest ϕ being larger when allowance is made for the systematic uncertainties.

In the standard profile likelihood, ν is a continuous variable. An extension of this has been used by Dauncey et al. [38], to allow for uncertainties in the choice of functional form of the background parametrisation in searches for new particles as peaks above background in a mass spectrum. Here the systematic is discrete, rather than continuous.

An alternative way of eliminating nuisance parameters (known as marginalisation) is to use $L(\phi, \nu)$ as part of a Bayesian procedure, and than to integrate the Bayesian posterior over ν. i.e.

$$P_{marg}(\phi) = \int P_{post}(\phi, \nu)\, d\nu \tag{15.5}$$

Of course, both profiling and marginalisation result in the loss of information. Reference [37] provides a very trivial example of this for profile likelihoods.

15.2.4 Punzi Effect

Sometimes we have two or more nearby peaks, and we try to fit our data in order to determine the fractions of each peak. Punzi [39] has pointed out that it is very easy to write down a plausible but incorrect likelihood function that gives a biassed result. This occurs in situations where the events have experimental resolutions σ in the observable x that vary event-by-event; and the distributions of σ are different for the two peaks.

For a set of observations x_i, it is tempting but wrong to write the unbinned likelihood as

$$L(f)_{wrong} = \Pi\{f * G(x_i, 0.0, \sigma_i) + (1 - f) * G(x_i, 1.0, \sigma_i)\} \tag{15.6}$$

where f is the fraction of the first peak (labelled A below) which is parametrised as $G(x_i, 0.0, \sigma_i)$, a Gaussian in x_i, centred on zero, and with width σ_i, and i is the label for the ith event; and similarly for the second peak (labelled B), except that it is centred at unity.

Application of the rules of conditional probability shows that the correct likelihood is

$$L(f)_{right} = \Pi\{f * G(x_i, 0.0, \sigma_i) * p(\sigma_i|A) + (1 - f) * G(x_i, 1.0, \sigma_i) * p(\sigma_i|B)\} \tag{15.7}$$

where $p(\sigma_i|A)$ and $p(\sigma_i|B)$ are the probability densities for the resolution being σ_i for the A and B peaks respectively. We then see that $L(f)_{wrong}$ and $L(f)_{right}$ give identical values for f, provided that $p(\sigma_i|A) = p(\sigma_i|B)$. If however, the distributions of the resolution differ, $L(f)_{wrong}$ will in general give a biassed estimate.

Punzi investigated the extent of this bias in a simple Monte Carlo simulation, and it turns out to be surprisingly large. For example, with $f = 1/3$, and $p(\sigma_A)$ and $p(\sigma_B)$ being $\delta-$functions at 1.0 and at 2.0 respectively (i.e. $\sigma = 1$ for all A events, and $\sigma = 2$ for all B events), the fitted value of f from $L(f)_{wrong}$ turned out to be 0.65. Given that f is confined to the range from zero to unity, this is an enormous bias.

The way the bias arises can be understood as follows: The fraction f of the events that are really A have relatively good resolution, and so the fit to them alone would assign essentially all of them as belonging to A i.e. these events alone would give $f \approx 1$ with a small uncertainty. In contrast the $1 - f$ of the events that are B have poor resolution, so for them the fit does not mind too much what is the value of f. But the fit uses all the events together, and so assigns a single f to the complete sample; this will be a weighted average of the f values for the A and for the B events. Because the A events result in a more accurate determination of f than do the B events, the fitted f will be biassed upwards (i.e. it will over-estimate the fraction of events corresponding to the peak with the better resolution).

The Punzi effect can also appear in other situations, such as particle identification. Different particle types (e.g. pions and kaons) would appear as different peaks in the relevant particle-identification variable e.g. time of flight, rate of energy loss dE/dx, angle of Cherenkov radiation, etc. The separation of these peaks for the different particle types depends on the momentum of the particles (see Fig. 15.2). The incorrect L is now

$$L_{wrong}(f_K) = \Pi\{(1-f_K)*G(x_i, x_\pi(p_i), \sigma_i)+f_K*G(x_i, x_K(p_i), \sigma_i)\} \tag{15.8}$$

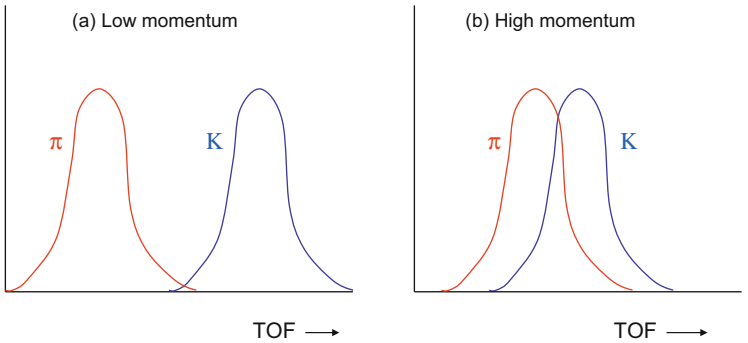


Fig. 15.2 The Punzi effect in particle identification. The diagrams show the expected (normalised) distributions of the output signal from a particle identifier, for pions and for kaons (**a**) at low momentum where separation is easier, and (**b**) at high momentum where the distributions overlap. Because kaons are heavier than pions, they tend to have larger momenta. Because it is hard at high momentum to distinguish pions from kaons, the likelihood function is insensitive to whether these tracks are classified as pions or kaons, and hence the fraction of high momentum tracks classified as kaons will have a large uncertainty. In contrast, low momentum tracks will be correctly identified. Thus if the plausible but incorrect likelihood function that ignores the pion and kaon momentum distributions is used to determine the overall fraction of kaons, it will be biassed downwards towards the fraction of low momentum particles that are kaons

where $x_\pi(p_i)$ and $x_K(p_i)$ are the expected positions of the particle identification information for a particle of momentum p_i, and x_i is the observed value for the ith event. So here the Punzi bias can arise even with constant resolution, because the momentum spectra of pions and kaons can be different. To avoid the bias, the likelihood needs to incorporate information on the different momentum distributions of pions and of kaons. If these momentum distributions are different enough from each other, it could be that the likelihood function bases its separation of the different particle types on the momenta of the particles rather than on the data from the detector's particle identifier. Catastini and Punzi[40] avoid this by using parametric forms for the momentum distributions of the particles, with the parameters being determined by the data being analysed.

The common feature potentially leading to bias in these two examples is that the ratio of peak separation to resolution is different for the two types of objects. For the first example of separating the two peaks, it was the denominators that were different, while in the particle identification problem it was the numerators.

The Punzi bias may thus occur in situations where the templates in a multi-component fit depend on additional observations whose distributions are not explicitly included in the likelihood.

15.3 Separating Signal from Background

Almost every Particle Physics analysis uses some technique for separating possible signal from background. First some simple 'cuts' are applied; these are generally loose selections on single variables, which are designed to remove a large fraction of the background while barely reducing the real or potential signal. Then to obtain a better separation of signal from background in the multi-dimensional space of the event observables, methods like Fisher discriminants, decision trees, artificial neural networks (including Bayesian nets and more recently deep neural nets), support vector machines, etc. are used[41, 42]. Extensions of these methods involve bagging, boosting and random forests, which have been used to achieve improved performance of the separation as seen on a plot of signal efficiency against background mis-acceptance rate. A description of the software available for implementing some of these techniques can be found in the talks by Narsky[43] and by Tegenfeldt[44] at the PHYSTAT-LHC Workshop.

More recently, deep learning techniques are rapidly becoming popular. In Particle Physics, they have been used for on-line triggering, tracking, fast simulation, object identification, image recognition, and event-by-event separation of signal from background. Reference [45] provides good introductions to the use of these methods for Particle Physics. There are now regular workshops and lectures on Machine Learning at CERN and at Fermilab (see refs. [46] and [47]), as well as at many universities.

The signal-to-background ratio before this multivariate stage can vary widely, as can the signal purity after it. If some large statistics study is being performed (e.g. to use a large sample of events to obtain an accurate measurement of the lifetime of some particle), then it is not a disaster if there is some level of background in the finally selected events, provided that it can be accurately assessed and allowed for in the subsequent analysis. At the other extreme, the separation technique may be used to see if there is any evidence for the existence of some hypothesised particle (the potential signal), in the presence of background from well-known sources. Then the actual data may in fact contain no observable signal.

These techniques are usually 'taught' to recognise signal and background by being given examples consisting of large numbers of events of each type. These may be produced by Monte Carlo simulation, but then there is a problem of trying to verify that the simulation is a sufficiently accurate representation of reality. It is better to use real data for this, but the difficulty then is to obtain sufficiently pure samples of background and signal. Indeed, for the search for a new particle, true data examples do not exist. However, it is the accurate representation of background that is likely to pose a more serious problem.

The way that, for example, neural networks are trained is to present the software with approximately equal numbers of signal and background events[10] and then

[10]For searches for rare processes, it is clearly inappropriate to use the actual fractions expected in the data to determine the ratio of signal to background Monte Carlo events to be used as the

to minimise a cost function C for the network. This is usually defined as $C = \Sigma(z_i - t_i)^2$, where z_i is the trained network's output for the ith event; t_i is the target output, usually chosen as 1 for signal and zero for background; and the summation is over all testing events presented to the network. The problem with this is that C is only loosely related to what we really want to optimise. For a search for a new particle this could be the sensitivity of the experimental upper limit in the absence of signal, while for a high statistics analysis measuring the properties (such as mass or lifetime) of some well-established particle, we would be interested in minimising the uncertainty (including systematic effects) on the result, without the training procedure biassing the measurement.

As with all event separation methods. it is essential to check the performance of a trained procedure by using a set of events that are independent of those used for training. This is to ensure that the network does not use specific but irrelevant features of the training events in its learning process, but can achieve good performance on unseen data.

Some open questions are:

- How can we check that our multi-dimensional training samples for signal and background are reliable descriptions of reality; and that they cover the region of multi-dimensional space populated by the data?
- How should the ratio of the numbers of signal and background training events be chosen, especially when there are several different sources of background?
- What is the best way of allowing for nuisance parameters in the models of the signal and/or background?[25, 48]
- Are there useful and easy ways of optimising on what is really of interest?[49]

15.3.1 Understanding How Neural Networks Operate

It is useful to appreciate how neural networks operate in providing a good separation of signal and background, as this can help in choosing a suitable architecture for the network.

Figure 15.3a shows some hypothetical signal and background events in terms of two measured variables x and y for each event. A network with two inputs (x and y), a single hidden layer with 3 nodes, and a single output is used; it aims to give 1 for signal and zero for background events (see Fig. 15.3b. This is achieved by training the network with (x, y) values for known examples of signal and background; and allowing the network to vary its internal parameters to minimise a suitably defined cost function e.g. $\Sigma(z_e - t_e)^2$, where the summation is over the training events, and z_e and t_e are the network's output and its target value (0 or 1) respectively.

training sample, because the network could then achieve a very small cost C simply by classifying everything as background.

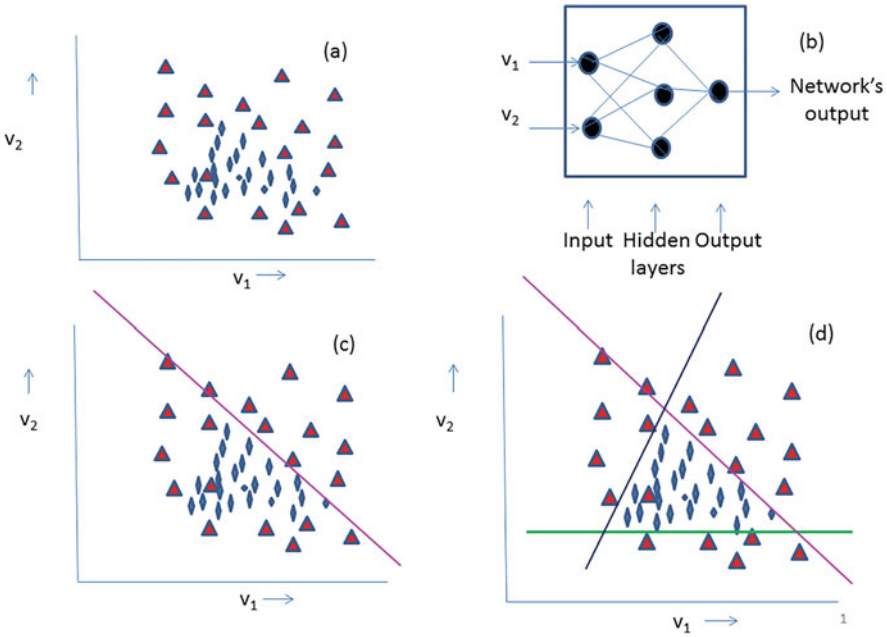


Fig. 15.3 (**a**) A $2-D$ plot showing the regions of the variables v_1 and v_2 for the signal (dots)and background (triangles). (**b**) The neural network used for separating signal and background. (**c**) The top hidden node receives inputs from v_1 and v_2. With suitable weights and threshold and a large value of β, the node's output will be on for (v_1, v_2) values below the diagonal line. (**d**) Similarly for the other two hidden nodes, their outputs can be on for $(v1, v2)$ below the other diagonal line, and above the horizontal one, respectively. A further choice of weights to the output node and its threshold can ensure that the whole network's output will be on only if all three hidden nodes' outputs are on, i.e. if (v_1, v_2) values are within the triangle in (**d**)

The input q_i to a given hidden node i is a linear combination of the input variables x and y

$$q_i = w_{xi}x + w_{yi}y + t_i \tag{15.9}$$

where the weights w and threshold t are varied during the fitting process. The output r from any hidden node is determined from its input q by something like a sigmoid function e.g.

$$r = 1/(1 + e^{-\beta q}) \tag{15.10}$$

This switches from zero for large negative q to unity for large positive q. The switch occurs around $q = 0$, and width of the region depends on the network parameter β. For very large β, there is a rapid switch from zero to unity. In terms of x and y, this

means that the hidden node i is 'on' (i.e. $r_i = 1$) if

$$w_{xi}x + w_{yi}y + t_i > 0, \tag{15.11}$$

or 'off' otherwise. Thus the boundary between events having r_i on or off is a straight line in the (x, y) plane (see Fig. 15.3c). With suitable values for the weights and thresholds for the three hidden nodes, there will be three straight line boundaries in the (x, y) plane shown in Fig. 15.3d. Finally, to produce the "and" of these three conditions, the weights w_{jo} (from the hidden node j to the output node o) and the output threshold t_o can be set as

$$w_{1o} = w_{2o} = w_{3o} = 0.4 \qquad t_o = -1.0 \tag{15.12}$$

to ensure that the output will be "on" only if the three hidden layers are all "on", i.e. that the selected input values are inside the triangular region in the $(x.y)$ plane. With β set at a lower level, the contour for the selected region will be smoother with rounded corners, rather than being triangular.

It would be useful to have a similar understanding of how deep networks operate. Tishby[50] has provided some insight on what happens in the hidden layers of a deep neural network during the training procedure.

15.4 Parameter Determination

For a single parameter (e.g. the branching ratio for $H \rightarrow \mu^+\mu^-$) the parameter range could be either a 2-sided interval or just an upper limit, at some confidence level (typically 68% for 2-sided intervals, but usually 90% and 95% for upper limits). For two parameters (e.g. mass and production rate for some new particle X that decays to a top pair), their acceptable values could be those inside some 2-dimensional confidence region. Alternatively an upper limit or 2-sided region for one parameter as a function of the other could be defined; these are known as a Raster Scan.

An upper limit on 2-variables is not a well-defined concept.

15.4.1 Upper Limits

Most recent searches for new phenomena have not found any evidence for exciting new physics. Examples from particle physics include searches for SUSY particles, dark matter, etc.; attempts to find substructure of quarks or leptons; looking for extra spatial dimensions; measuring the mass of the lightest neutrino; etc. Rather than just saying that nothing was found, it is more useful to quote an upper limit on the sought-for effect, as this could be useful in ruling out some theories. For example in

1887, Michelson and Morley[52] attempted to measure the speed of the Earth with respect to the aether. No effect was seen, but the experiment was sensitive enough to lead to the demise of the aether theory.

A simple scenario is a counting experiment where a background b is expected from conventional sources, together with the possibility of an interesting signal s. The number of counts n observed is expected to be Poisson distributed with a mean $\mu = \epsilon s + b$, where b is the expected number of events from background, and ϵ is a factor for converting the basic physics parameter s into the number of signal events expected in our particular experiment; it thus allows for experimental inefficiency, the experiment's running time; etc. Then given a value of n which is comparable to the expected background, what can we say about s? The true value of the parameter s is constrained to be non-negative. The problem is interesting enough if b and ϵ are known exactly; it becomes more complicated when only estimates with uncertainties σ_b and σ_ϵ are available.

An extension of the simple counting scenario is when a search for a new particle is carried out over a range of masses. This is usually dealt with by performing separate searches at a series of masses over a specified range. This 'Raster Scan' is in contrast with a method that regards the sought-for new particle's mass and its production rate as two parameters to be estimated simultaneously. The relative merits of these two approaches are described in ref. [51].

Even without the nuisance parameters, a variety of methods is available. These include likelihood, χ^2, Bayesian with various priors for s, frequentist Neyman constructions with a variety of ordering rules for n, and various *ad hoc* approaches. The methods give different upper limits for the same data.[11] A comparison of several methods can be found in ref. [53]. The largest discrepancies arise when the observed n is less than the expected background b, presumably because of a downward statistical fluctuation. The following different behaviours of the limit (when $n < b$) can be obtained:

- Frequentist methods can give **empty** intervals for s i.e. there are no values of s for which the data are likely. Particle physicists tend to be unhappy when their years of work result in an empty interval for the parameter of interest, and it is little consolation to hear that frequentist statisticians are satisfied with this feature, as it does not necessarily lead to undercoverage.
  When n is not quite small enough to result in an empty interval, the upper limit might be **very small**.[12] This could confuse people into thinking that the experiment was much more sensitive than it really was.
- The Feldman-Cousins frequentist method[54] (see Sect. 15.4.3) that employs a likelihood-ratio ordering rule gives upper limits which **decrease** as n gets smaller

[11] By coincidence, the upper limits obtained by the Bayesian approach with an (improper) flat prior for s and by the appropriate Neyman construction agree when $b = 0$.

[12] Bayesian methods that use priors with part of the probability density being a δ-function at $s = 0$ can result is a posterior with an enhanced δ-function at zero, such that the upper limit contains only the single point $s = 0$.

at constant b. A related effect is the growth of the limit as b decreases at constant n—this can also occur in other frequentist approaches. Thus if no events are observed ($n = 0$), the upper limit of a 90% Feldman-Cousins interval is 1.08 for $b = 3.0$, but 2.44 for $b = 0$. This is sometimes presented as a paradox, in that if a bright graduate student worked hard and discovered how to eliminate the expected background without much reduction in signal efficiency, the 'reward' would be a weaker upper limit.[13] An answer is that although the actual limit had increased, the sensitivity of the experiment with the smaller background was better. There are other situations—for example, variants of the random choice of voltmeter (compare ref. [55])—where a measurement with better sensitivity can on occasion give a less precise result.

- In the Bayesian approach, the dependence of the limit on b is **weaker**. Indeed when $n = 0$, the limit does not depend on b.
- Sen et al. [56] consider a related problem, of a physical non-negative parameter λ producing a measurement x, which is distributed about λ as a Gaussian of variance σ^2. As the observable x becomes more and more negative, the upper limit on λ **increases**, because it is deduced that σ must in fact be larger than its quoted value.

In trying to assess which of the methods is best, one first needs a list of desirable properties. These include:

- Coverage: Even though coverage is a frequentist concept, most Bayesian particle physicists would like the coverage of their intervals to match their reported credibility, at least approximately.
  Because the data in counting experiments is discrete, it is impossible in any sensible way to achieve exact coverage for all μ (see Sect. 15.1.3.4). However, it is not completely obvious that even Frequentists need coverage for every possible value of μ, since different experiments will have different values of b and of ϵ. Thus even for a constant value of the physical parameter s, different experiments will have different $\mu = \epsilon * s + b$. Thus it would appear that, if coverage in some average (over μ) sense were satisfactory, the frequentist requirement for intervals to contain the true value at the requisite rate would be maintained. This, however, is not the generally accepted view by particle physicists, who would like not to undercover for **any** μ.
- Not too much overcoverage: Because coverage varies with μ, for methods that aim not to undercover anywhere, some overcoverage is inevitable. This corresponds to having some upper limits which are high, and this leads to undesirable loss of power in rejecting alternative hypotheses about the parameter's value.

[13] The $n = 0$ situation is perhaps a special case, as the number of observed events cannot decrease as further selections are imposed to reduce the expected background. For non-zero observed events, if n decreases with the tighter cuts (as expected for reduced background), the upper limit is likely to go down, in agreement with intuition. But if n stays constant, that could be because the observed events contain signal, so it is perhaps not surprising that the upper limit increases.

- Short and empty intervals: These can be obtained for certain values of the observable, without resulting in undercoverage. They are generally regarded as undesirable for the reasons explained above.

It is not obvious how to incorporate the above desiderata on interval length into an algorithm that would be useful for choosing among different methods for setting limits. For different experiments studying the same phenomena (e.g. Dark Matter searches, neutrino oscillation experiments, etc.) it is worthwhile to use the same technique for calculating allowed parameter ranges.

15.4.2 Two-Sided Intervals

An alternative to giving upper limits is to quote two-sided intervals. For example, a 68% confidence interval for the mass of the top quark might be 172.6–173.4 GeV/c^2, as opposed to its 95% upper limit being 173.6 GeV/c^2. Most of the difficulties and ambiguities mentioned above apply in this case too, together with some extra possibilities. Thus, while it is clear which of two possible upper limits is tighter, this is not necessarily so for two-sided intervals, where which is shorter may be metric dependent; the first of two intervals for a particle's lifetime τ may be shorter, but the second may be shorter when the ranges are quoted for its decay rate ($= 1/\tau$). There is also scope for choice of ordering rule for the frequentist Neyman construction, or for choosing the interval from the Bayesian posterior probability density.[14]

15.4.3 Feldman-Cousins Approach

Feldman and Cousins' fully frequentist approach[54] exploits the freedom available in the Neyman construction of how to choose an interval in the data that contains a given fraction α of the probability, by using their 'ordering rule'. This is based on the likelihood ratio $L(x, \mu)/L(x, \mu_{best})$, where μ_{best} is the physically-allowed value of μ which gives the largest value of L for that particular x. For values of μ far from a physical boundary, this makes little difference from the standard central Neyman construction, but near a boundary the region is altered in such a way as to make it unlikely that there will be zero-length or empty intervals for the parameter μ; these can occur in the standard Neyman construction (see Fig. 15.4).

[14] A Bayesian statistician would be happy with the posterior as the final result. Particle physicists like to quote an interval as a convenient summary. For a parameter that cannot be negative and for which the exclusion of zero is interesting (e.g. testing whether the production rate of some hypothesised particle is non-zero), an upper limit would always include zero, a lower limit or a central interval would exclude it and a maximum probability density one would not be invariant with respect to changes in the functional form of the parameter.

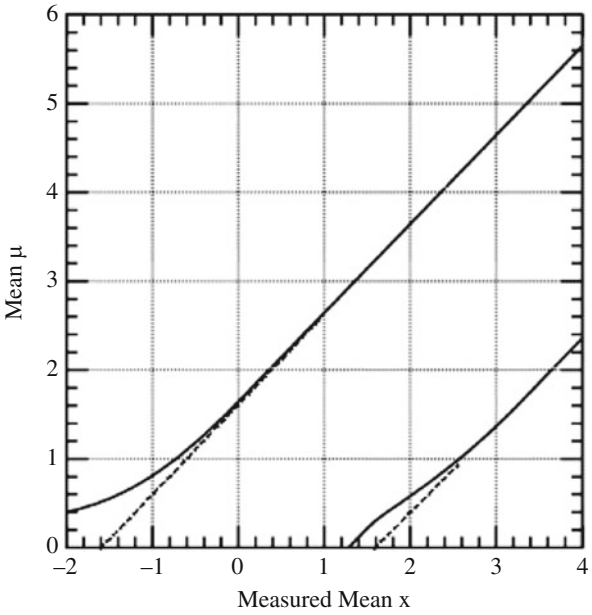


Fig. 15.4 The Feldman-Cousins 90% confidence band (solid curves) for the mean μ of a Gaussian probability density function of unit variance for a measurement x. The straight dashed lines show the confidence band for the central Neyman construction. The Feldman-Cousins ordering rule pulls the interval to the left at small μ, and hence, even for negative observed x, the μ interval is not empty, as happens for central frequentist intervals when x is below -1.6

The original Feldman-Cousins paper also considered how to extend their method when there is more than one parameter and one measurement. They describe an idealised neutrino oscillation experiment with the data being the energy spectrum of the interacting neutrinos, and the parameters are $\sin^2(2\theta)$ and Δm^2 (see Eq. 15.15). A practical problem of having many parameters is the CPU time required to compute the results.

Feldman and Cousins also point out that an apparently innocuous procedure for choosing what result to quote may lead to undercoverage. Many physicists would quote an upper limit on any possible signal if their observation was less than 3 standard deviations above the expected background, but a two-sided interval if their result was above this. With each type of interval constructed to give 90% coverage, there are some values of the parameter for which the coverage for this mixed procedure drops to 85%; Feldman and Cousins refer to this as 'flip-flop'. Their 'unified' approach circumvents this problem, as it automatically yields upper limits for small values of the data, but two-sided intervals for larger measurements, while avoiding undercoverage for all possible true values of the signal.

15.4.4 Sensitivity

It is useful to quote the sensitivity of a procedure, as well as the actual upper limit as derived from the observed data.[15] For upper limits or for uncertainties on measurements, this can be defined as the median value that would be obtained if the procedure was repeated a large number of times.[16] Using the median is preferable to the mean because (a) it is metric independent (i.e. the median lifetime upper limit would be the reciprocal of the median decay rate lower limit); and (b) it is much less sensitive to a few anomalously large upper limits or uncertainty estimates.

It is common to present not only the median of the expected distribution, but also values corresponding to 16th and 84th percentiles (commonly referred to as $\pm 1\ \sigma$) and also the 2.5% and 97.5% ones ($\pm 2\sigma$). This enables a check to be made that the observed result is reasonable.

Punzi [57] has drawn attention to the fact that this choice of definition for sensitivity has some undesirable features. Thus designing an analysis procedure to minimise the median upper limit for a search in the absence of a signal provides a different optimisation from maximising the median number of standard deviations for the significance of a discovery when the signal is present. Also there is only a 50% chance of achieving the median result or better. Instead, for pre-defined levels α and confidence level CL, Punzi determines at what signal strength there is a probability of at least CL for establishing a discovery at a significance level α. This is what he quotes as the sensitivity, and is the signal strength at which we are sure to be able either to claim a discovery or to exclude its existence. Below this, the presence or otherwise of a signal makes too little difference, and we may remain uncertain (see Fig. 15.5).

15.4.5 Nuisance Parameters

For calculating upper limits in the simple counting experiment described in Sect. 15.4.1, the nuisance parameters arise from the uncertainties in the background rate b and the acceptance ϵ. These uncertainties are usually quoted as σ_b and σ_ϵ (e.g. $b = 3.1 \pm 0.5$), and the question arises of what these uncertainties mean. Sometimes they encapsulate the results of a subsidiary measurement, performed to estimate b or ϵ, and then they would express the width of the Bayesian posterior or of the frequentist interval obtained for the nuisance parameters. However, in

[15] The sensitivity on its own will not do, because it is independent of the data.

[16] Instead of using a large number of simulations in order to extract the median, sometimes the 'Asimov' data set is used. This is the single data set that would be obtained if statistical fluctuations were suppressed. i.e. if a model predicted 11.3 events in a particular bin, the Asimov data set for that model would contain 11.3 events in that bin. The Asimov data set and the median of the toys usually but not always produce similar results.

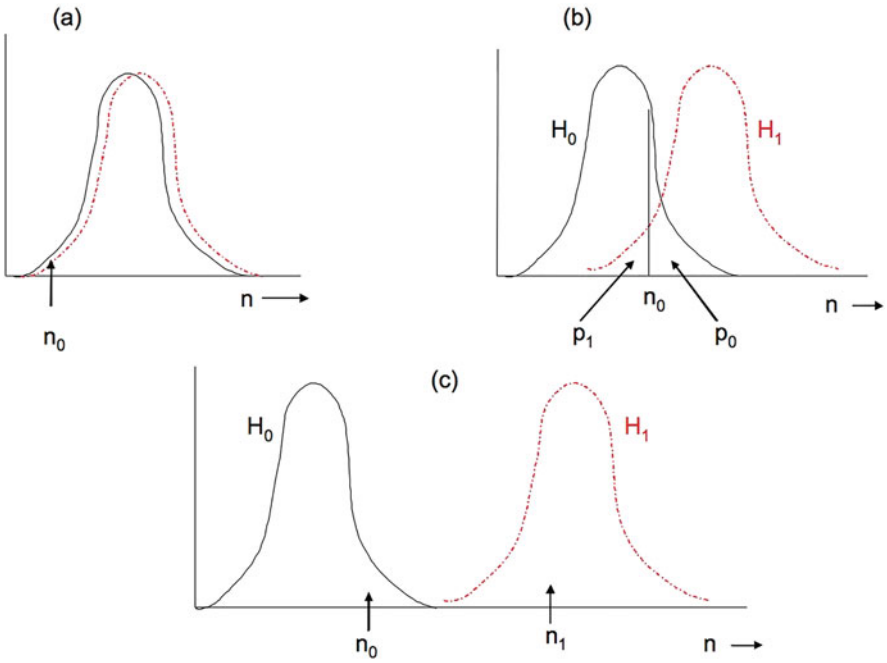


Fig. 15.5 Punzi definition of sensitivity. Expected distributions for a statistic t (which in simple cases could be simply the observed number of events n), for H_0 = background only (solid curves) and for H_1 = background plus signal (dashed curves). In (**a**), the signal strength is very weak, and it is impossible to choose between H_0 and H_1. As shown in (**b**), which is for moderate signal strength, p_0 is the probability according to H_0 of t being equal to or larger than the observed t_0. To claim a discovery, p_0 should be smaller than some pre-set level α, usually taken to correspond to 5σ; t_{crit} is the minimum value of t for this to be so. Similarly p_1 is the probability according to H_1 for $t \leq t_0$. The power function is the probability according to the alternative hypothesis that t will exceed t_{crit}. As the separation of the H_0 and H_1 pdfs increases, so does the power. According to Punzi, the sensitivity should be defined as the expected production strength of the signal such that the power exceeds another predefined CL, e.g. 95%. The exclusion region corresponds to t_0 in the 5% lower tail of H_1, while the discovery region has t_0 in the 5σ upper tail of H_0; in (**b**) there is a "No decision" region in between, as the signal strength is below the sensitivity value. The sensitivity is thus the signal strength above which there is a 95% chance of making a 5σ discovery. i.e. The distributions for H_0 and H_1 are sufficiently separated that, apart possibly for the 5σ upper tail of H_0 and the 5% lower tail of H_1, they do not overlap. In (**c**) the signal strength is so large that there is no ambiguity in choosing between the hypotheses

many situations, the uncertainties may involve Monte Carlo simulations, which have systematic uncertainties (e.g. related to how well the simulation describes the real data) as well as statistical ones; or they may reflect uncertainties or ambiguities in theoretical calculations required to derive b and/or ϵ. In the absence of further information the posterior is often assumed to be a Gaussian, usually truncated so as to exclude unphysical (e.g. negative) values. This may be at best only approximately

true, and deviations are likely to be most serious in the tails of the distribution. A log-normal or gamma function may be a better choice.

There are many methods for incorporating nuisance parameters in upper limit calculations. These include:

- Profile likelihood (see also Sect. 15.2.3)
  The likelihood, based on the data from the main and from the subsidiary measurements, is a function of the parameter of interest s and of the nuisance parameters. The profile likelihood $L_{prof}(s)$ is simply the full likelihood $L(s, b_{best}(s), \epsilon_{best}(s))$, evaluated at the values of the nuisance parameters that maximise the likelihood at each s. Then the profile likelihood is simply used to extract the limits on s, much as the ordinary likelihood could be used for the case when there are no nuisance parameters.
  Rolke et al. [59] have studied the behaviour of the profile likelihood method for limits. Heinrich[32] had shown that the likelihood approach for estimating a Poisson parameter (in the absence of both background and of nuisance parameters) can have poor coverage at low values of the Poisson parameter. However, the profile likelihood seems to do better, probably because the nuisance parameters have the effect of smoothing away the fluctuating coverage observed by Heinrich.
- Fully Bayesian
  When there is a subsidiary measurement for a nuisance parameter, a prior is chosen for b (or ϵ), the data are used to extract the likelihood, and then Bayes' Theorem is used to deduce the posterior for the nuisance parameter. This posterior from the subsidiary measurement is then used as the prior for the nuisance parameter in the main measurement (this prior could alternatively come from information other than a subsidiary measurement); with the prior for s and the likelihood for the main measurement, the overall joint posterior for s and the nuisance parameter(s) is derived.[17] This is then integrated over the nuisance parameter(s) to determine the posterior for s, from which an upper limit can be derived; this procedure is known as marginalisation.
  Numerical examples of upper limits can be found in ref. [60], where a method is discussed in detail. Thus assuming (somewhat unrealistically) precisely determined backgrounds, the effect of a 10% uncertainty in ϵ can be seen for various measured values of n in Table 15.1. A plot of the coverage when the uncertainty in ϵ is 20% is reproduced in Fig. 15.6.
  It is not universally appreciated that the choice for the main measurement of a truncated Gaussian prior for ϵ and an (improper) constant prior for non-negative s results in a posterior for s which diverges[61]. Thus numerical estimates of the relevant integrals are meaningless. Another problem comes from the difficulty of choosing sensible multi-dimensional priors. Heinrich has pointed out the

[17] This is usually equivalent to starting with a prior for s and the nuisance parameters, and the likelihood for the data from the main and the subsidiary experiments together, to obtain the joint posterior.

Table 15.1 Bayesian 90% confidence level upper limits for the production rate s as a function of n, the observed number of events

n	$b = 0.0$	$b = 3.0$
0	2.35 (2.30)	2.35 (2.30)
3	6.87 (6.68)	4.46 (4.36)
6	10.88 (10.53)	7.80 (7.60)
9	14.71 (14.21)	11.56 (11.21)
20	28.27 (27.05)	25.05 (24.05)

The Poisson parameter $\mu = \epsilon * s + b$, where the expected background b is either 0.0 or 3.0, and is precisely known; and ϵ, whose true values is 1.0, is estimated in a subsidiary measurement with 10% accuracy. The numbers in brackets are the corresponding upper limits when ϵ is known precisely. At large n, the limits for $b = 3.0$ are 3 units lower than those for $b = 0.0$; the latter are approximately $n + 1.28\sqrt{n}$ at large n. The effect of the uncertainty in ϵ is to increase the limits, and by a larger amount at large n. For $n = 0$, these Bayesian limits are independent of the expected background b

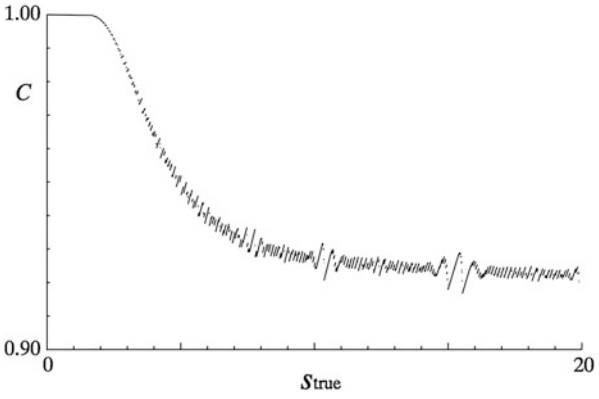


Fig. 15.6 The coverage C for the 90% confidence level upper limit as a function of the true parameter s_{true}, as obtained in a Bayesian approach. The background $b = 3.0$ is assumed to be known exactly, while the subsidiary measurement for ϵ gives a 20% accuracy. The discontinuities are a result of the discrete (integer) nature of the measurements. There is no undercoverage

problems that can arise for the above Poisson counting experiment, when it is extended to deal with several data channels simultaneously[62].

- Fully frequentist
  In principle, the fully frequentist approach to setting limits when provided with data from the main and from subsidiary measurements is straightforward: the Neyman construction is performed in the multidimensional space where the parameters are s and the nuisance parameters, and the data are from all the relevant measurements. Then the region in parameter space for which the observed data was likely is projected onto the s-axis, to obtain the confidence region for s.

In practice there are formidable difficulties in writing a program to do this in a reasonable amount of time. Another problem is that, unless a clever ordering rule is used for producing the acceptance region in data space for fixed values of the parameters, the projection phase leads to overcoverage, which can become larger as the number of nuisance parameters increases. Good ordering rules have been found for a version of the Poisson counting experiment[63], and also for the ratio of Poisson means[64], where the confidence intervals are tighter than those obtained by conditioning on the sum of the numbers of counts in the two observations.

For the fully frequentist method, it is guaranteed that there will be no undercoverage for any combination of parameter true values. This is not so for any other method, and so most particle physicists would like assurance that the technique used does indeed provide reasonable coverage, at least for s. There is usually lively debate between frequentists and Bayesians as to whether coverage is desirable for all values of the nuisance parameter(s), or whether one should be happy with no or little undercoverage when experiments are averaged, for example, over the nuisance parameter true values.

- Mixed

  Because of the difficulty of performing a fully frequentist analysis in all but the simplest problems, an alternative approach[65] is to use Bayesian averaging over the nuisance parameters, but then to employ a frequentist approach for s. The hope is that for most experiments setting upper limits, the statistical uncertainties on the low n data are relatively large and so, provided the uncertainties in the nuisance parameters are not too large, the effect of the systematics on the upper limits will not be too dramatic, and an approximate method of dealing with them may be reasonable.

  Although such an approach cannot be justified from fundamentals, it provides a practical method whose properties can be checked, and is often satisfactory.

15.4.6 Banff Challenges

Given the large number of techniques available for extracting upper limits from data, especially in the presence of nuisance parameters, it was decided at the Banff meeting[6] that it would be useful to compare the properties of the different approaches under comparable conditions. This led to the setting up of the 'Banff Challenge', which consisted of providing common data sets for anyone to calculate their upper limits. This was organised by Joel Heinrich, who reported on the performance of the various methods at the PHYSTAT-LHC meeting[66].

At the second Banff meeting[8], the challenge was set by Tom Junk and consisted of participants trying to distinguish between histograms, some of which contained only background and others which contained a background and signal, which appeared as a peak (compare Sect. 15.7.5)

15.4.7 Recommendations

It would be incorrect to say that there is one method that must be used. Many Particle Physicists' ideal would be to use a frequentist approach if viable software were available for problems with several parameters and items of data. Otherwise they would be prepared to settle for a Bayesian approach, with studies of the sensitivity of the upper limit to the choice of priors, and of the coverage; or for a profile likelihood method, again with coverage studies. What is important is that the procedure should be fully defined before the data are analysed; and that when the experimental result and the sensitivity of the search are reported, the method used should be fully explained.

The CDF Statistics Committee [67] also suggests that it is useful to use a technique that has been employed by other experiments studying the same phenomenon; this makes for easier comparison. They tend to favour a Bayesian approach, chiefly because of the ease of incorporating nuisance parameters.

15.5 Combining Results

This section deals with the combination of the results from two or more measurements of a single (or several) parameters of interest. It is not possible to combine upper limits (UL). This is because an 84% UL of 1.5 could come from a measurement of 1.4 ± 0.1, or 0.5 ± 1.0; these would give very different results when combined with some other measurement.

The combination of p-values is discussed in Sect. 15.7.9.

15.5.1 Single Parameter

An interesting question is whether it is possible to combine two measurements of a single quantity, each with uncertainty ± 10, such that the uncertainty on the combined best estimate is ± 1? The answer can be deduced later.

To combine N different uncorrelated measurements $a_i \pm \sigma_i$ of the same physical quantity a[18] when the measurements are believed to be Gaussian distributed about the true value a_{true}, the well-known result is that the best estimate $a_{comb} \pm \sigma_{comb}$ is given by

$$a_{comb} = \Sigma(a_i * w_i)/\Sigma w_i, \quad \sigma_{comb} = 1/\sqrt{\Sigma w_i}, \tag{15.13}$$

where the weights are defined as $w_i = 1/\sigma_i^2$. This is readily derived from minimising with respect to a a weighted sum of squared deviations

$$S(a) = \Sigma(a_i - a)^2/\sigma_i^2 \tag{15.14}$$

The extension to the case where the individual measurements are correlated (as is often the case for analyses using different techniques on the same data) is straightforward: $S(a)$ becomes $\Sigma\Sigma(a_i - a) * H_{ij} * (a_j - a)$, where H is the inverse covariance matrix for the a_i. It provides **B**est **L**inear **U**nbiassed **E**stimates (BLUE)[70].

There are, however, practical details that complicate its application. For example, in the above formula, the σ_i are supposed to be the **true** accuracies of the measurements. Often, all that we have available are **estimates** of their values. Problems arise in situations where the uncertainty estimate depends on the measured value a_i. For example, in counting experiments with Poisson statistics, it is typical to set the uncertainty as the square root of the observed number. Then a downward fluctuation in the observation results in an overestimated weight, and a_{comb} is biassed downwards. If instead the uncertainty is estimated as the square root of the expected number a, the combined result is biassed upwards—the increased uncertainty reduces S at larger a. A way round this difficulty has been suggested by Lyons et al. [71]. Alternatively, for Poisson counting data a likelihood approach is preferable to a χ^2-based method.

Another problem arises when the individual measurements are very correlated. When the correlation coefficient of two uncertainties is larger than σ_1/σ_2 (where σ_1 is the smaller uncertainty), a_{comb} lies outside the range of the two measurements. As the correlation coefficient tends to +1, the extrapolation becomes larger, and is sensitive to the exact values assumed for the elements of the covariance matrix. The situation is aggravated by the fact that σ_{comb} tends to zero. This is usually dealt with by selecting one of the two analyses, rather than trying to combine them. However, if the estimated uncertainty increases with the estimated value, choosing the result with the smaller **estimated** uncertainty can again produce a downward bias. On the other hand, using the smaller **expected** uncertainty can cause us to ignore an analysis which had a particularly favourable statistical fluctuation, which produced a result

[18] It is of course much better to use all the **data** in a combined analysis, rather than simply to combine the **results**.

that was genuinely more precise than expected[19] How to deal with this situation in general is an open question. It has features in common with the problem (inspired by ref. [55]) of measuring a voltage by choosing at random a voltmeter from a cupboard containing meters of different sensitivities.

Another example involves combining two measurements of a cross-section with small statistical uncertainties, but with large correlated uncertainties from the common luminosity. With this luminosity uncertainty included in the covariance matrix, BLUE can result in the combined value being outside the range of the individual measurements. For this situation, it is preferable to exclude the luminosity uncertainty from the covariance matrix, and to apply it to the combined result afterwards.

15.5.2 Two or More Parameters

An extension of this procedure is for combining N pairs of correlated measurements (e.g. the gradient and intercept of a straight line fit to several sets of data, where for simplicity it is assumed that any pair is independent of every other pair). For several pairs of values (a_i, b_i) with inverse covariance matrices $\mathbf{M}_i$, the best combined values (a_{comb}, b_{comb}) have as their inverse covariance matrix $\mathbf{M} = \Sigma\mathbf{M}_i$. This means that, if the covariance matrix correlation coefficients ρ_i of the different measurements are very different from each other, the uncertainty on a_{comb} can be much smaller than that for any single measurement.

This situation applies for track fitting to hits in a series of groups of tracking chambers, where each set of close chambers provides a very poor determination of the track; but the combination involves widely spaced chambers and determines the track well. Using the profile likelihoods (e.g. for the intercept, profiled over the gradient) for combining different measurements loses the correlation information and can lead to a very poor combined estimate[37]. The alternative of ignoring the correlation information is also strongly discouraged.

The importance of retaining covariances is relevant for many combinations, e.g. for the determination of the amount of Dark Energy in the Universe from various cosmological data[73].

[19]For example, the ALEPH experiment at LEP produced a tighter-than-expected upper limit on the mass of ν_τ because they happened to observe τ decay configurations which were particularly sensitive to the ν_τ mass.

15.5.3 Data Consistency

The standard procedure for combining data pays no attention to whether or not the data are consistent. If they are clearly inconsistent, then they should not all be combined. When they are somewhat inconsistent, the procedure adopted by the Particle Data Group[14] is to increase all the uncertainties by a common factor such that the overall χ^2 per degree of freedom equals unity.[20]

The Particle Data Group prescription for expanding uncertainties in the case of discrepant data sets has complications when each of the data sets consists of two or more parameters[72].

15.6 Goodness of Fit

15.6.1 Sparse Multi-Dimensional Data

The standard method loved by most scientists uses the weighted sum of squares, commonly called χ^2. This, however, is only applicable to binned data (i.e. in a one or more dimensional histogram). Furthermore it loses its attractive feature that its distribution is model-independent when there is not enough data, which is likely to be so in the multi-dimensional case.

Although the maximum likelihood method is very useful for parameter determination with **unbinned data**, the value of L_{max} usually does not provide a measure of goodness of fit (see Sect. 15.2.2).

An alternative that is used for sparse one-dimensional data is the Kolmogorov-Smirnov (KS) approach[68], or one of its variants. However, in the presence of fitted parameters, simulation is again required to determine the expected distribution of the KS-distance. Also because of the problem of how to order the data, the way to use it in multi-dimensional situations is not unique.

The standard KS method uses the maximum deviation between two cumulative distributions; because of statistical fluctuations, this is likely to occur near the middle of the distributions. In cases where interesting New Physics is expected to occur at extreme values of some kinematic variable (e.g. p_T), variants of KS such as Anderson-Darling[69] that give extra weight to the distributions' tails may be more useful.

[20]This is somewhat conservative, in that even if there are no problems, about half the data sets would be expected to have this larger than unity.

15.6.2 Number of Degrees of Freedom

If we construct the weighted sum of squares S between a predicted theoretical curve and some data in the form of a histogram, provided the Poisson distribution of the bin contents can be approximated by a Gaussian (and the theory is correct, the data are unbiased, the uncertainty estimates are correct, etc.), **asymptotically**[21] S will be distributed as χ^2 with the number of degrees of freedom $\nu = n - f$, where n is the number of data points and f is the number of free parameters whose values are determined by minimising S.

The relevance of the asymptotic requirement can be seen by imagining fitting a more or less flat distribution by the expression $N(1 + 10^{-6}\cos(x - x_0))$, where the free parameters are the normalisation N and the phase x_0. It is clear that, although x_0 is left free in the fit, because of the 10^{-6} factor, it will have a negligible effect on the fitted curve, and hence will not result in the typical reduction in S associated with having an extra free parameter. Of course, with an enormous amount of data, we would have sensitivity to x_0, and so asymptotically it does reduce ν by one unit, but not for smaller amounts of data.

Another example involves neutrino oscillation experiments[54]. In a simplified two neutrino scenario, the neutrino energy spectrum is fitted by a survival probability P of the form

$$P = 1 - \sin^2 2\theta \ \sin^2(C * \Delta m^2), \tag{15.15}$$

where C is a known function of the neutrino energy and the length of its flight path, Δm^2 is the difference in mass squared of the relevant neutrino species, and θ is the neutrino mixing angle. For small values of $C * \Delta m^2$, this reduces to

$$P \approx 1 - \sin^2 2\theta \ (C * \Delta m^2)^2 \tag{15.16}$$

Thus the survival probability depends on the two parameters only via their product $\sin 2\theta \, \Delta m^2$. Because this combination is all that we can hope to determine, we effectively have only one free parameter rather than two. Of course, an enormous amount of data can manage to distinguish between $\sin\left(C * \Delta m^2\right)$ and $C * \Delta m^2$, and so asymptotically we have two free parameters as expected.

15.7 Discovery Issues

Searches for new particles are an exciting endeavour, and continue to play a large role at the LHC at CERN, in neutrino experiments, in searches for dark matter, etc. The 2007 and 2011 PHYSTAT Workshops at CERN[7, 9] were devoted specifically

[21] The examples in this section go beyond the requirement that we need enough events for the Poisson distribution to be well approximated by a Gaussian.

to statistical issues that arise in discovery-orientated analyses at the LHC. Ref [74] deals with statistical issues that occur in Particle Physics searches for new phenomena; as an example, it includes the successful search for the Higgs boson at the LHC. A more detailed description of the plans for the Higgs search before its discovery is in ref. [75].

15.7.1 H_0, or H_0 Versus H_1?

In looking for new physics, there are two distinct types of approach. We can compare our data just with the null hypothesis H_0, the SM of Particle Physics; alternatively we can see whether our data are more consistent with H_0 or with an alternative hypothesis H_1, some specific manifestation of new physics, such as a particular form of quark and/or lepton substructure. The former is known as 'goodness of fit', while the term 'hypothesis testing' is often reserved for the latter.

Each of these approaches has its own advantage. By not specifying a specific alternative,[22] the goodness of fit test may be capable of detecting any form of deviation from the SM. On the other hand, if we are searching for some specific new effect, a comparison of H_0 and H_1 is likely to be a more sensitive way for that particular alternative. Also, the 'hypothesis testing' approach is less likely to give a false discovery claim if the assumed form of H_0 has been slightly mis-modelled.

15.7.2 p-Values

In order to quantify the chance of the observed effect being due to an uninteresting statistical fluctuation, some statistic is chosen for the data. The simplest case would be the observed number n_0 of interesting events. Then the p-value is calculated, which is simply the probability that, given the expected background rate b from known sources, the observed value would fluctuate up to n_0 or larger. In more complicated examples involving several relevant observables, the data statistic may be a likelihood ratio L_0/L_1 for the likelihood of the null hypothesis H_0 compared with that for a specific alternative H_1.

To compute the p-value of the observed or of possible data, the distribution $f(t)$ of the data statistic t under the relevant hypothesis is required. In some cases this can be obtained analytically, but in more complicated situations, $f(t)$ may require simulation. For t being $-2\ln L_0/L_{best}$, Cowan et al have given useful asymptotic

[22] Even a test of the null hypothesis may not be completely independent of ideas about alternatives. Thus in an event counting experiment, new physics usually results in an **increase** in rate, unless we are looking for neutrino oscillations, in which case a **decrease** would be significant. Also, sometimes the statistic used for a goodness of fit test of H_0 may be the likelihood ratio for H_0 as compared with a specific alternative H_1.

formulae for $f(t)$[76]; here L_{best} is the value of the likelihood when the parameters in H_0 are set at their best values.

A small value of p indicates that the data are not very compatible with the theory (which may be because the detector's response or the background is poorly modeled, rather than the theory being wrong).

Particle Physicists usually convert p into the number of standard deviations σ of a Gaussian distribution, beyond which the one-sided tail area corresponds to p; statisticians refer to this as the z-score, but physicists call it significance. Thus 5σ corresponds to a p-value of $3 * 10^{-7}$. This is done simply because it provides a number which is easier to remember, and not because Gaussians are relevant for every situation.

Unfortunately, p-values are often misinterpreted as the probability of the theory being true, given the data. It sometimes helps colleagues clarify the difference between $p(A|B)$ and $p(B|A)$ by reminding them that the probability of being pregnant, given the fact that you are female, is considerably smaller than the probability of being female, given the fact that you are pregnant. Reference [77] contains a series of articles by statisticians on the use (and misuse) of p-values.

Sometimes $S/\sqrt{B}$ or $S/\sqrt{(S+B)}$ or the like (where S is the number of observed events above the estimated background B) is used as an approximate measure of significance. These approximations can be very poor, and their use is in general not recommended.[23]

15.7.3 CL_s

This is a technique[58] which is used for situations in which a discovery is not made, and instead various parameter values are excluded. For example the failure to observe SUSY particles can be converted into mass ranges which are excluded (at some confidence level).

Figure 15.5 (again) illustrates the expected distributions for some suitably chosen statistic t under two different hypotheses: the null H_0 in which there is only standard known physics, and H_1 which also includes some specific new particle, such as a SUSY neutralino. In Fig. 15.5c, the new particle is produced prolifically, and an experimental observation of t should fall in one peak or the other, and easily distinguishes between the two hypotheses. In contrast, Fig. 15.5a corresponds to very weak production of the new particle and it is almost impossible to know whether the new particle is being produced or not.

[23]For example, if selections to enhance signal with respect to background were optimised using $S/\sqrt{B}$, extremely hard cuts might be chosen, yielding expected numbers of events $S = 0.1$ and $B = 10^{-3}$. This results in $S/\sqrt{B} = 10$, which sounds very good, but in fact this selection is disastrous.

The conventional method of claiming new particle production would be if the observed t fell well above the main peak of the H_0 distribution; typically a p_0 value corresponding to 5σ would be required (see Sect. 15.7.7). In a similar way, new particle production would be excluded if t were below the main part of the H_1 distribution. Typically a 95% exclusion region would be chosen (i.e. $p_1 \leq 0.05$), where p_1 is by convention the left-hand tail of the H_1 distribution, as shown in Fig. 15.5b.

The CL_s method aims to provide protection against a downward fluctuation of t in Fig. 15.5a resulting in a claim of exclusion in a situation where the experiment has no sensitivity to the production of the new particle; this could happen in 5% of experiments. It achieves this by defining[24]

$$CL_s = p_1/(1 - p_0), \tag{15.17}$$

and requiring CL_s to be below 0.05. From its definition, it is clear that CL_s cannot be smaller than p_1, and hence is a conservative version of the frequentist quantity p_1. It tends to p_1 when t lies above the H_0 distribution, and to unity when the H_0 and H_1 distributions are very similar. The reduced CL_s exclusion region is shown by the dotted diagonal line in Fig. 15.7; the price to pay for the protection provided by CL_s is that there is built-in conservatism when p_1 is small but p_0 has intermediate values i.e. there are more cases in which no decision is made. Most statisticians are appalled by the use of CL_s, because they consider that it is meaningless to take the ratio of two p-values.

It is deemed not to be necessary to protect against statistical fluctuations giving rise to discovery claims in situations with no sensitivity, because that should happen only at the $3 * 10^{-7}$ rate (the one-sided 5σ Gaussian tail area).

Figure 15.7 is also useful for understanding the Punzi sensitivity definition (see Sect. 15.4.4). For any specified distributions of the statistic t for H_0 and H_1, the possible (p_0, p_1) values lie on a curve or straight line which extends from (0,1) to (1,0). With more data, the t distributions separate, and the curve moves closer to the p_0 and p_1 axes. The amount of data required to satisfy the Punzi requirement of always claiming a discovery or an exclusion is when no part of the curve is in the "no decision" region of Fig. 15.7.

[24]Given the fact that CL_s is the ratio of two p-values, the choice of symbol CL_s (standing for 'confidence level of signal') is not optimal. Another source of confusion is that in definitions of CL_s the ways the p-values are defined vary, so the formulae can look different but the underlying concept is the same.

A subtlety with Eq. (15.17) is that p_0 there is the probability of obtaining a measurement **greater than** the observed one, rather than the usual 'greater than or equal to'. This is to make $1 - p_0$ the probability of a value smaller than or equal to the observed one, in analogy with the definition of p_1. It makes a difference when the observation is a small discrete number.

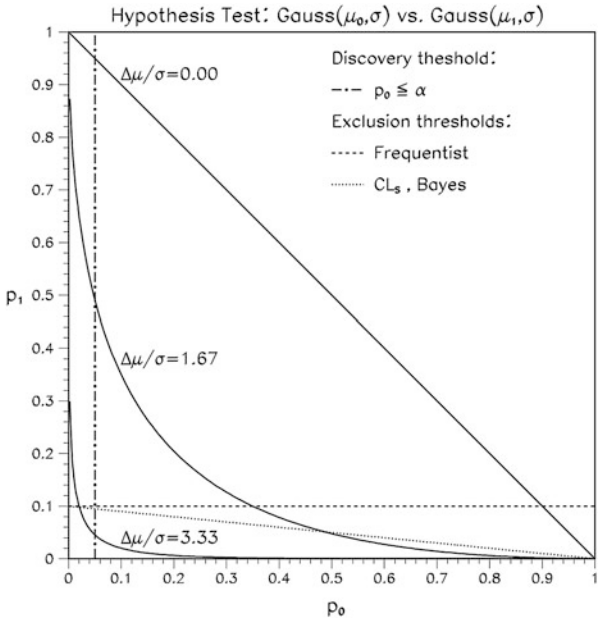


Fig. 15.7 Plot of p_0 against p_1 for comparing a data statistic t with two hypotheses H_0 and H_1, whose expected pdf's for t are given by two Gaussians of peak separation $\Delta\mu$, and of equal width σ. For a given pair of pdf's for t, the allowed values of (p_0, p_1) lie on a curve or straight line (shown solid in the diagram). The expected density for the data along a curve is such that its projection along the p_0-axis (or p_1-axis) is expected to be uniform for the hypothesis H_0 (or H_1 respectively). As the separation increases, the curves approach the p_0 and p_1 axes. Rejection of H_0 is for p_0 less than, say, $3 * 10^{-7}$; here it is shown as 0.05 for ease of visualisation. Similarly exclusion of H_1 is shown as $p_1 < 0.1$. Thus the (p_0, p_1) square is divided into four regions: the largest rectangle is when there is no decision, the long one above the p_0-axis is for exclusion of H_1, the high one beside the p_1-axis is for rejection of H_0, and the smallest rectangle is when the data lie between the two pdf's. For $\Delta\mu/\sigma = 3.33$, there are no values of (p_0, p_1) in the "no decision" region. In the CL_s procedure, rejection of H_1 is when the t statistic is such that (p_0, p_1) lies below the diagonal dotted straight line

15.7.4 Comparing Two Hypotheses Via χ^2

Assume that there is a histogram with 100 bins, and that a χ^2 method is being used for fitting it with a function with one free parameter. The expected value of χ^2 is 99 ± 14. Thus if p_0, the best value of the parameter, yields a χ^2 of 85, this would be regarded as very satisfactory. However, a theoretical colleague has a model which predicts that the parameter should have a different value p_1, and wants to know what the data have to say about that. This is tested by calculating the χ^2 for that p_1, which yields a value of 110. There appear to be two contradictory conclusions:

- p_1 is satisfactory: This is based on the fact that the relevant χ^2 of 110 is well within the expected range of 99 ± 14.

- p_1 is ruled out: The uncertainty on p is estimated by seeing how much it must change from its optimum value in order to make χ^2 increase by **1**. For this data, $\chi^2(p_1)$ is **25** units larger than $\chi^2(p_0)$, and so, assuming that the behaviour of χ^2 in the neighbourhood of the minimum is parabolic, p_1 is ruled out at the $\sim$5 standard deviation level.

Unfortunately, many physicists, over-impressed by the fact that $\chi^2(p_1)$ appears to be satisfactory, are reluctant to accept that p_0 is strongly favoured by the data.

A similar argument applies to comparing a given set of data with 2 separate hypotheses e.g. fitting a histogram with an exponential or a straight line. Again the **difference** between the χ^2 quantities provides better discrimination between the hypotheses than do the **individual** χ^2 values[78]. Another example of using the difference in χ^2's is given in the next section.

There are of course other ways available for comparing two hypotheses. e.g. likelihood ratio, Bayes factor, Bayesian information criterion, etc. For a fuller discussion, see ref. [79]. A description of their application in cosmology can be found in ref. [80]. Problems in choosing priors for the Bayes factor approach for selecting among hypotheses are discussed by Heinrich[81].

15.7.5 Peak Above Smooth Background

When comparing two hypotheses with our data, we can use the numerical values of the two χ^2 quantities with a view to making some decision about the hypotheses. For example, we may be fitting a smooth distribution by a power series, and wonder whether we need a quadratic term, or whether a linear expression would suffice. Alternatively we may want to assess whether a mass spectrum favours the existence of a peak on top of a smooth background, as compared with just the smooth background. Qualitatively, if the extra term(s) are unnecessary, they will result in a relatively small reduction in χ^2, while if they really are required, the reduction could be larger.

It is sometimes possible to be quantitative about the expected reduction when the extra terms are not needed[82]. If we are in the asymptotic regime, and if the hypotheses are nested,[25] and if the extra parameters of the larger hypothesis are defined under the smaller one, and in that case do not lie on the boundary of their allowed region, then the difference in χ^2 should itself be distributed as a χ^2, with the number of degrees of freedom equal to the number of extra parameters.

An example that satisfies this is provided by the different order polynomials. The hypotheses are nested, in that the linear situation is a special case of a quadratic, where the coefficient of the quadratic term is zero. Thus the extra parameter is defined and within the (infinite) allowed range. Then, provided we have a large

[25]This means that for suitable values of the parameters the larger hypothesis reduces to the smaller one.

amount of data, we expect the difference in χ^2 to have one degree of freedom, so a value larger than around 5 would be unlikely.

A contrast is provided by a smooth background $C(x)$ compared with a background plus peak, $C(x) + A \exp[-0.5 * (x - x_0)^2/\sigma^2]$. The extra parameters for the peak are its amplitude, position and width: A, x_0 and σ respectively. Again the hypotheses are nested, in that $C(x)$ is just a special case of the peak plus background, with $A = 0$. However, although A is defined in the background only case, x_0 and σ are not, as their values become completely irrelevant when $A = 0$. Furthermore, unless the peak plus background fit allows A to be negative, zero is on the boundary of its allowed region. We thus should not expect the difference of the χ^2 quantities itself to be distributed as a χ^2 [83–85]. To assess the significance of a particular χ^2 difference, this unfortunately means that we have to obtain its distribution ourselves, presumably by Monte Carlo. If we want to find out probabilities of statistical fluctuations at the 10^{-6} level, this requires a lot of simulation, and probably needs us to use something better than brute force.

The problem of non-standard limiting distributions for χ^2 tests has a substantial statistical literature (see, for example, refs. [86] and [87].)

15.7.6 Incorporating Nuisance Parameters

The calculation of p-values is complicated in practice by the existence of nuisance parameters. (For the simple situation described in Sect. 15.7.2, there could be some uncertainty in the estimated background.) There are numerous ways of incorporating them. These include:

- Conditioning: For example, with a single nuisance parameter, it may be possible to condition on the sum of the number of counts in the main and the subsidiary experiments, and then to use the binomial distribution to obtain the p-value.
- Plug-in p-value: The best estimate of the nuisance parameter under the null hypothesis is used to calculate p.
- Prior predictive p-value: The p-values are averaged over the nuisance parameters, weighted by their prior distributions. This is in the spirit of the Cousins and Highland approach[65] for upper limits.
- Posterior predictive p-value: This time, the posterior distributions of the nuisance parameters are used for weighting.
- Supremum p-value: The largest p-value for any possible value of the nuisance parameter is used. This is likely to be useful only when the nuisance parameter is forced to be within some range; or when there is only a small number of possible alternative theoretical interpretations.
- Confidence interval: A region of frequentist confidence $1 - \gamma$ is used for the nuisance parameter(s), and then the adjusted p-value is $p_{max} + \gamma$, where p_{max} is the largest p-value as the nuisance parameters are varied over their confidence

region. Clearly if it is desired to establish a discovery from p-values around 10^{-7} or smaller, then γ should be chosen at least an order of magnitude below this.

The properties of these and other methods are compared by Demortier [84], while Cranmer [88] and Cousins et al.[89] have discussed some of them in the context of searches at the LHC.

The role of systematic effects is likely to be more serious here than for upper limits discussed in Sect. 15.4.5. This is because in upper limit situations the number of events is usually small, and so statistical uncertainties dominate. In contrast, discovery claims have p-values of $3 * 10^{-7}$ or smaller, and so tails of distributions are likely to be important.

15.7.7 Why 5σ?

Unfortunately the usually accepted criterion for claiming a discovery in Particle Physics is that p should correspond to at least 5σ. Statisticians almost invariably ask why such a stringent level is used. One answer is past experience: all too often interesting effects at the 3σ or 4σ level have gone away as more data are collected. Another is the multiple comparison problem, or "Look Elsewhere Effect" (LEE). While the chance of obtaining a 5σ effect in one bin of a particular histogram ("local p-value") is really small, it is to be remembered that histograms have many bins,[26] they could be plotted with different selection criteria and different binning,[27] and there are very many other histograms that were or could have been looked at in the course of the experiment.[28] Thus the chance of a 5σ fluctuation occurring somewhere in the data ("global p-value") is much larger than might at first appear. Calculating a global p-value may require an excessive amount of Monte-Carlo simulation. Reference [90] circumvents this for asymptotic situations by providing a formula for extrapolating the LEE correction factor from a lower significance level; this requires considerably less simulation.

Finally, physicists subconsciously incorporate Bayesian priors in assessing how likely they feel that they have discovered something new, and hence whether they

[26] In calculating a p-value in such a case, it is very desirable to take into account the number of chances for a statistical fluctuation to occur anywhere in the histogram (or anywhere in the search procedure, for more complicated analyses). At very least, it should be made clear what the basis of the calculated p-value is.

[27] If a blind analysis is performed, such decisions are made before looking at the data, and so this aspect of the "look elsewhere" effect is reduced.

[28] The extent to which other people's searches should be included in an allowance for the "look elsewhere" effect depends on the implied question being addressed. Thus are we considering the chance of obtaining a statistical fluctuation in any of the analyses we have performed; or by anyone analysing data in our experiment; or by any Particle Physicist this year? Because of the ambiguity of which specific question is being addressed, which is often not explicitly mentioned, we recommend not including an extra "look elsewhere" factor for this.

should claim a discovery. Thus, in deciding between the possibilities of a new discovery or of an undetected systematic effect, our priors might favour the latter, and hence strong evidence for discovery is required from the data.[29]

However it is not necessarily equitable to use a uniform standard for large general-purpose experiments and for small ones with a specific aim; or for looking for a process which is expected (e.g. $H^0 \rightarrow \mu^+\mu^-$), as compared with a more speculative search, such as lepton substructure[91]. But physicists and especially journal editors seem to like a defined rule rather than a flexible criterion, so this bolsters the 5σ standard. In any case, it is largely a semantic issue, in that physicists finding a 4.5σ effect would clearly report it, using judiciously chosen wording to describe the interpretation of their observation.

Statisticians also ask whether models can really be trusted to describe the extreme tails of distributions. In general, this may be so—counting experiments are expected to follow Poisson distributions, with small corrections for possible long time-scale drifts in detector calibrations; and particle decays usually are described by exponential distributions in time. However, the situation is much less clear for nuisance parameters, where uncertainty estimates may be less rigorous, and their distribution is often assumed to be Gaussian (or truncated Gaussian) by default. The effect of these uncertainties on very small p-values needs to be investigated case-by-case.

It is important to remember that p-values merely test the null hypothesis. There are more sensitive ways of looking for new physics when a specific alternative is relevant. Thus a very small p-value on its own is usually not enough to make a convincing case for discovery.

15.7.8 Repetitions in Time

Often experiments accumulate data over several years. The same search for a new effect may typically be repeated once or twice each year as more data are collected. Does this constitute another factor of $\sim$20 in the number of opportunities for a statistical fluctuation to appear? Our reply is "No". If there had been a 6σ signal with the early data (which resulted in a claim for discovery), which had then become only 3σ with more data, this would be grounds for downplaying the earlier discovery claim. Thus at any time, there is essentially only one set of data (everything) that is relevant.

For a p-value to be meaningful, it is important that the time at which the experiment stops collecting data is determined not by the significance of the observed signal but by external factors (e.g. accelerator being decommissioned, ending of funding, etc.). Indeed there is a theorem that states that, provided data is

[29] If I were performing an experiment to look for violations of energy conservation, I would require more than 5σ, because my prior for energy being conserved is very large.

collected for long enough, it is possible to reach any arbitrary level of significance against a hypothesis that is in fact true.

15.7.9 Combining p-Values

In looking for a given new effect, there may be several separate and uncorrelated analyses which are relevant. These could correspond to different decay modes for the new particle; or different experiments looking for the same signal. Thus, if the p-values for the null hypothesis (i.e. no new physics) for the separate analyses were 10^{-6} and 0.1, what is the corresponding p-value for the pair of results?[30]

The unambiguous answer is that there is no unique recipe for combining them[92, 93]. There is no single way of taking a uniform distribution in two variables, and finding a transformation $p_{comb}(p_1, p_2)$ that converts it into a uniform distribution of the single variable p_{comb}.

Two popular recipes involve asking what is the probability that the smaller p-value will be 10^{-6} or smaller; or that the product is below $p_1 * p_2 = 10^{-7}$. (Note that these probabilities are **not** 10^{-6} and 10^{-7} respectively.) None of the possible methods has the property that in combining three p-values, the same answer is obtained if p_1 is first combined with p_2, and then the result is combined with p_3; or whether some different ordering is used.

Another problem is the lack of other information that might be relevant. For example, the p-values might arise from χ^2's with different numbers of degrees of freedom ν e.g. $\chi^2_1 = 90$ for 100 degrees of freedom, and $\chi^2_2 = 20$ for $\nu = 1$. The second has a very small p-value, so many combination methods (including the two mentioned above) would conclude that overall the data do not look consistent with the null hypothesis. However, another plausible-sounding method is to add the separate χ^2 values and also the individual ν,[31] to obtain a total $\chi^2 = 110$ for $\nu = 101$, which sounds perfectly satisfactory. The resolution of this discrepancy of interpretation depends on the nature of the two tests. If the second analysis with $\chi^2 = 20$ corresponded to just one extra measurement like the previous 100, then it seems reasonable to combine the χ^2 values and the ν, and to conclude that overall there is indeed nothing surprising. But on the other hand, if the second measurement was genuinely different, and an alternative way of looking for some discrepancy, then it may be more appropriate to combine the p-values by one of the earlier methods, which suggest that the overall consistency with theory is not good. It

[30] Rather than combining p-values, it is of course much better to use the complete sets of original data (if available) for obtaining the combined result.

[31] The method described earlier involving the product of the p-values is equivalent to converting each p to a χ^2, assuming that $\nu = 2$, regardless of whether this was the actual number of degrees of freedom, and then adding the χ^2 and also the ν.

is this extra information about the nature of the two tests that determines which combination method might be appropriate.

It is clearly important to decide in advance what combination method should be used, without reference to the specific data being analysed.

15.8 Blind Analyses

These are becoming increasingly popular as a means of avoiding personal bias affecting the result. They involve keeping part of the data unseen by the analysers, until the data selection procedure and the analysis method have been completely defined, all correction procedures specified, etc.

One of the early suggestions to use a blind analysis in a Particle Physics experiment was due to Luis Alvarez. An experiment at Stanford had looked for quarks, by measuring the residual charge on small spheres that were levitated in a superconducting magnet. If a single free quark were present in a sphere, the residual charge would be a third or two-thirds of the electron's charge. Several of the balls tested indeed yielded such values[94]. A potential problem was that large corrections had to be applied to the raw data in order to extract the final result for the charge. The suspicion was that maybe the experimenters were (subconsciously) applying corrections until the value turned out to be 'satisfactory'. The blind approach involved the computer adding a random number to the raw value of the charge, which would then be corrected until the experimentalists were satisfied, and only then would the computer subtract the random number to reveal the final answer for that sphere.[32]

There are various methods of performing blind analyses[95] most of which aim to allow the experimentalists to look at some of the real data, in order to perform checks that nothing is terribly wrong. Some of these are:

- The computer adds a random number to the data, which is only subtracted after all corrections are applied. This was the method suggested by Alvarez.
- Use only Monte Carlo to define the procedure. This completely avoids the danger of allowing the data to determine the procedure to be used, but suffers from the drawback that the data cannot be compared with the Monte Carlo, to check that the latter is reasonable.
- Use only a fraction of the data for defining the procedure, which then is held fixed for the remainder of the data. In principle, an optimisation can be employed to determine the fraction to be kept open, but in practice this is often decided by choosing a semi-arbitrary time after which the future data is kept blind.

[32]This suggestion was implemented, but in fact no subsequent results were published. The current consensus is that this 'discovery' of free quarks is probably spurious.

- The signal region is defined by a certain part of multi-dimensional space, and this is kept hidden, but all other regions, including those adjacent to the signal, are available for inspection.
- Keep the Monte Carlo parameters hidden. This is a technique suggested by the TWIST experiment in their high statistics precision determination of parameters associated with muon decay. The procedure involves comparing the data with various simulated sets, generated with a series of different parameter values. The data and the simulations are both visible, but the parameter values used to generate the simulations are kept hidden.
- Keep visible only a fraction of the contents of each bin of a histogram. This is used by the MINOS experiment searching for neutrino oscillations; these would affect the energy distribution of the observed events. By keeping visible different unknown fractions of the data in each bin, the energy spectral shape cannot be determined from the visible part of the data.

If several different groups within the same collaboration are performing similar analyses for extracting some specific parameter, then it is desirable to fix the procedure for selecting which result to present, or alternatively how to combine the separate results. This should be done before the results are seen, and is worth doing even if the individual analyses were not "blind".

A question that arises with blind analyses is whether it should be permitted to modify the analysis after the data had been unblinded. It is generally agreed that this should not be done, unless everyone would regard it as ridiculous not to do so. For example, if a search for rare events yielded 10 candidates over the course of a year's run, all of which occurred on Sunday mornings at precisely 1.17 a.m., it would be prudent to do some further investigation before publishing. If 'post-unblinding' modification of the procedure is performed, this should be made clear in any publication.

15.9 Topics that Deserve More Attention

15.9.1 Statistical Software

Particle physicists tend to write their own software for performing statistical computations. Although this has educational merits, it is inefficient use of one's time. The data-manipulation system of programmes ROOT/RooFit/RooStats contains many useful statistical routines[96]. Tools also exist for implementing many methods for separating signal from background[43, 44].

A problem with these is that they are too easy to use. In the hands of a non-critical user, the required input data instructions may contain some error, with the consequence that they will produce the solution to a different procedure than the intended one. It is very important to check that the result obtained is not unreasonable.

15.9.2 Deep Learning

This involves the use of sophisticated techniques[45] for achieving nearly optimal extraction of information from data, but which are still relatively unfamiliar to many scientists. It is important to develop a set of protocols to ensure that they perform in a reliable manner, and are not introducing subtle biases of which users are unaware.

15.9.3 Unfolding Data or Smearing Theory?

Observed experimental distributions are almost always smeared versions of 'the true distributions of Nature'. It is simpler to compare theory and data by smearing the theory, rather than trying to unfold the experimental effects from the data, as the latter is a less stable procedure and also introduces correlations among the bins of the unfolded distribution. Some fields tend to favour deconvolution; this is partly because it is rarer for them to have a dominant theoretical model with which the data is to be compared. Unfolding does have the advantage that it provides an estimate of the 'true' distribution, with which any future theory can be compared. Also it can be looked at by a physicist, but we are not accustomed to readily interpreting data where the contents of the histogram bins are highly correlated.

There are some situations where unfolding is desirable. For example, it allows the comparison of distributions from different experiments, with different resolutions. Another is using experimental data for tuning Monte Carlo generators; smearing the data at each step of the optimisation increases the computation time too much.

Even for checking in future whether new theories are compatible with data does not necessarily require unfolding. Provided that the smearing matrix of the detector is provided, the future data can be smeared, and then compared with the actual (not unfolded) data. However, including the effects of systematics can be a complication.

Sessions at the 2011 PHYSTAT workshop[9] and at CERN's PHYSTATν meeting[12] were devoted to unfolding. Blobel[97] has reviewed the topic, while ref. [98] contains a statistician's view of the statistical issues involved in unfolding.

15.9.4 Visualisation

The combination of the human eye and brain is very powerful at detecting patterns in data (even if sometimes they are not there!) This can be useful in deciding how to analyse the data; as a check on whether the result of an analysis is plausible; whether a machine learning method for separating signal from background is performing sensibly; etc. Such human inspection of data is feasible if there are only a small number (below 4) of relevant variables. Techniques for inspecting multi-dimensional data would be valuable.

15.9.5 Non-parametric Methods

These are so unknown to most Particle Physicists that they are usually unaware when they are using them. Simple examples include:

- A histogramme as an estimate of the density distribution of a variable of interest.
- Kernel density estimation.
- Kolmogorov-Smirnov or Anderson-Darling methods, to test whether distributions are consistent.
- Classification schemes based on k nearest neighbours.
- Neural networks

These all avoid the need to specify a particular parametric form, and hence the values of any parameters. In general such a method is less powerful than a parametric one, if the latter were available and relevant.

15.9.6 Collaboration with Statisticians

Other scientists seem to be better than particle physicists about involving statisticians in the analysis of their data. This is partly due to the fact that we like to try out statistical techniques ourselves; that we consider our data is too complicated for other people to deal with; and that we are somewhat over-protective of our data, and are reluctant to share it with others. None of this is particularly convincing, and it is clear that we would benefit from the involvement of professional statisticians. The advantages of having them participating in the recent PHYSTAT meetings have been obvious.

In the past, Particle Physicists have on occasion asked rather specific questions to Statisticians they happened to know. Statisticians prefer to be much more directly involved with the data itself. With analyses becoming more and more complex, it will be highly desirable for them to be affiliated with experimental groups.

15.10 Conclusion

Although the statistical aspects of many particle physics analyses are already at a sophisticated level, it is clear that there are many practical statistical issues to be resolved. With the increasing complexity of scientific investigations, more active collaboration with statisticians and machine learning experts will result in a better understanding of the relevant techniques and improved analyses in the future.

Acknowledgements I wish to acknowledge the patience and expertise of David Cox, Brad Efron, Jerry Friedman and David van Dyk, and also of other Statisticians too numerous to list, in explaining statistical issues to me; the ones who have contributed to the PHYSTAT meetings have

been particularly helpful. My understanding of the practical application of statistical techniques has improved considerably as a result of discussions with many experimental Particle Physics colleagues, and in particular with the members of the CDF and CMS Statistics Committees. I especially wish to thank Bob Cousins, Luc Demortier and Joel Heinrich for their careful reading and valuable comments on the original version of this article. To all of you, I am most grateful.

The Leverhulme Foundation kindly provided a grant which partially supported the original version this work.

References

1. Workshop on Confidence Limits, CERN Yellow Report 2000-05.
2. FNAL Confidence Limits Workshop (2000), http://conferences.fnal.gov/CLW/.
3. Advanced Statistical Techniques in Particle Physics, Durham (2002) IPPP/02/39.
4. Proceedings of PHYSTAT2003, eConf C030908, SLAC-R-703.
5. "PHYSTAT05: Statistical Problems in Particle Physics, Astrophysics and Cosmology", Imperial College Press (2006), http://www.physics.ox.ac.uk/phystat05/.
6. BIRS Workshop on "Statistical inference Problems in High Energy Physics and Astronomy", Banff (2006), http://www.birs.ca/birspages.php?task=displayevent&event_id=06w5054.
7. PHYSTAT-LHC Workshop on "Statistical Issues for LHC Physics" (2007), http://phystat-lhc.web.cern.ch/phystat-lhc/2008-001.pdf.
8. BIRS Workshop on "Statistical issues relevant to significance of discovery claims (10w5068)" (2010) https://www.birs.ca/events/2010/5-day-workshops/10w5068
9. PHYSTAT-LHC Workshop, "Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding", https://cdsweb.cern.ch/record/1306523/files/CERN-2011-006.pdf
10. PHYSTATν in Japan (2016), https://indico.cern.ch/event/735431/
11. PHYSTATν Workshop on Statistical issues in experimental neutrino physics, FNAL (2016), https://indico.fnal.gov/event/11906/
12. PHYSTATν in CERN (2019), https://indico.cern.ch/event/735431/
13. Roger Barlow, "Statistics: a Guide to the Use of Statistical Methods in the Physical Sciences", Wiley (1989).
    O. Behnke at al (eds), "Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods", Wiley (2013),
    Glen Cowan, "Statistical Data Analysis", Oxford University Press (1998).
    F. E. James, "Statistical Methods in Experimental Physics", World Scientific Publishing Co (2007).
    L. Lista, "Statistical Methods for Data Analysis in Particle Physics", Springer (2017).
    Louis Lyons, "Statistics for Nuclear and Particle Physics", Cambridge University Press (1986). See also https://www-cdf.fnal.gov/physics/statistics/notes/Errata2.pdf for an Update.
    Byron Roe, "Probability and Statistics in Experimental Physics", Springer Verlag (1991).
14. M. Tamabashi et al., "Review of Particle Physics", Phys. Rev. **D**98 030001 (2018).
15. BaBar Statistics Working Group, http://www.slac.stanford.edu/BFROOT/www/Statistics/.
16. CDF Statistics Committee, http://www-cdf.fnal.gov/physics/statistics/statistics_home.html
17. ATLAS Statistics Forum, https://twiki.cern.ch/twiki/bin/view/Atlas/StatisticsTools#Statistics_Forum.
18. CMS Statistics Committee, https://twiki.cern.ch/twiki/bin/view/CMS/StatisticsCommittee.
19. D. van Dyk, "Statistical quantification of discovery in neutrino physics", Neutrino2016 XXVII Int Conf on Neutrino Physics and Astrophysics, http://neutrino2016.iopconfs.org/home
20. L. Lyons, "Lessons learned from PhysStat-nu", NuPhys2016: Prospects in Neutrino Physics, https://indico.ph.qmul.ac.uk/indico/conferenceDisplay.py?confId=170; and

"Statistical issues towards PHYSTATν 2019", NuPhys2018: Prospects in Neutrino Physics, https://indico.ph.qmul.ac.uk/indico/conferenceDisplay.py?confId=289
21. CERN European Schools, https://physicschool.web.cern.ch/physicschool/ESHEP/previous_eshep.html;
CERN Latin-American Schools, https://physicschool.web.cern.ch/physicschool/CLASHEP/previous_clashep.html; and
CERN Asia-Pacific Schools, http://aepshep.org/previous-schools.html.
22. L. Lyons, "Statistical Issues in Particle Physics" in "Elementary Particles: Detectors for Particles and Radiation, Part 1: Principles and Methods" Eds C. Fabjan and H. Schopper, (Landolt-Bornstein, **21B1** 2011).
23. ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", Phys. Lett. B**716** (2012) 1;
CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", Phys. Lett. B**716** (2012) 30.
24. N. Reid, "Some aspects of design of experiments", ref. [7], p. 94.
25. R. Neal, "Computing likelihood functions when distributions are defined by simulations with nuisance parameters", ref. [7], p. 101; and in ref. [6].
26. J. Linnemann, "A pitfall in estimating systematic errors", ref. [7], p. 94.
27. J. Heinrich and L. Lyons, Annual Reviews of Nuclear and Particle Science **57** (2007) 145.
28. L. Lyons, "Bayes and Frequentism: a Particle Physicist's perspective", (2013) https://arxiv.org/pdf/1301.1273.pdf
29. R. D. Cousins, Am. J. Phys. **63** (1995) 398.
30. R. Barlow, "Asymmetric errors", ref. [4], p. 250; and ref. [5], p. 56.
31. F. Garwood, "Fiducial limits for Poisson the distribution", Biometrica **28** (1936) 437.
32. J. Heinrich, "Coverage of error bars for Poisson data" (2003), http://www-cdf.fnal.gov/publications/cdf6438_coverage.pdf.
33. J. Heinrich, "Pitfalls of Goodness-of-Fit from Likelihood", ref. [4], p. 52.
34. S. Baker and R. D. Cousins, "Clarification of the use of χ^2 and likelihood functions in fits to histograms", NIM **221** issue 2 (1984) 437.
35. G. Cowan, Eur Phys J C (2019) 79:133.
36. D. Cox, private communication
37. L. Lyons and E. Chapon, "Combining parameter values or p-values" (2017) https://arxiv.org/pdf/1704.05540.pdf
38. P. Dauncey et al, "Handling uncertainties in background shapes: the discrete profiling method", JINST **10** no.04 (2015) 04015
39. G. Punzi, "Comments on likelihood fits with variable resolution", ref. [4], p. 235.
40. P. Catastini and G. Punzi, "Bias-free estimation of multicomponent maximum likelihood fits with component-dependent templates", ref. [5], p. 60.
41. H. B. Prosper, "Multivariate methods: a unified perspective", ref. [3], p. 91.
42. J. H. Friedman, "Recent advances in predictive (machine) learning", ref. [4], p. 196; and "Separating signal from background using ensembles of rules", ref. [5], p. 127.
43. I. Narsky, "StatPatternRecognition in analysis of HEP and Astrophysics data", ref. [7], p. 188.
44. A. Hocker et al, "TMVA, Toolkit for Multi-Variate data Analysis with ROOT", ref. [7], p. 184.
45. D. Guest, K. Cranmer and D. Whiteson, "Deep Learning and Its Application to LHC Physics", Annual Review of Nuclear and Particle Science **68** (2018) 161;
A. Radovic et al, "Machine learning at the energy and intensity frontiers of particle physics", Nature **560** (2018) 41;
A. J. Larkoski, I. Moult and B. Nachman "Deep Learning and Its Application to LHC Physics" (2017) http://arxiv.org/abs/arXiv:1709.04464;
I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning" (20116), MIT Press.
46. CERN's "Inter-Experimental LHC Machine Learning Working Group (IML)", https://iml.web.cern.ch/
47. Fermilab's "ML at the Intensity and Cosmic Frontiers", https://machinelearning.fnal.gov/
48. L. Lyons, Nucl. Inst. Meth. **A324** (1993) 565.

49. S. Whiteson and D. Whiteson, "Stochastic Optimization for Collision Selection in High Energy Physics." IAAI 2007: Proceedings of the Nineteenth Annual Innovative Applications of Artificial Intelligence Conference (July 2007) 1819.
50. N. Tishby and N. Zaslavsky, "Deep Learning and the Information Bottleneck Principle" (2015), https://arxiv.org/abs/1503.02406;
R. Shwartz-Ziv and N. Tishby, "Opening the Black Box of Deep Neural Networks via Information" (2017), https://arxiv.org/abs/1703.00810
51. L. Lyons, "Raster scan or 2-D approach?", https://arxiv.org/pdf/1404.7395.pdf
52. A. Michelson and E.Morley, "On the Relative Motion of the Earth and the Luminiferous Ether", American Journal of Science. **34** (1887) 203: 333.
53. I. Narsky, "Comparison of upper limits", in ref. [2].
54. G. J. Feldman and R. D. Cousins, Phys. Rev. **D57** (1998) 3873.
55. D. R. Cox, "Some problems connected with statistical inference", Annals of Mathematical Statistics **29** (1958) 357.
56. B. Sen, M. Walker and M. Woodroofe, "On the Unified Method with Nuisance Parameters", Statistica Sinica **19** (2009) 301
57. G. Punzi, "Sensitivity of searches for new signals and its optimisation", ref. [4], p. 235.
58. A. L. Read, "Modified frequentist analysis of search results", ref. [1], p. 81; "Presentation of search results—the CL_s method", ref. [3], p. 11.
T. Junk, "Confidence level computation for combining searches with small statistics", NIM A **434** (1999) 435.
59. W. A. Rolke, A. M. Lopez and J. Conrad, Nuclear Instruments and Methods **A551** (2005) 493.
60. J. Heinrich et al. "Interval estimation in the presence of nuisance parameters. 1. Bayesian approach", CDF note 7117 (2004), https://www-cdf.fnal.gov/physics/statistics/notes/cdf7117_bayesianlimit.pdf
61. L. Demortier, "A fully Bayesian computation of upper limits for Poisson processes", CDF note 5928 (2004).
62. J. Heinrich, "The Bayesian approach to setting limits: what to avoid", ref. [5], p 98.
63. G. Punzi, "Ordering algorithms and confidence intervals in the presence of nuisance parameters", ref. [5], p. 88.
64. R. Cousins, Nuclear Instruments and Methods **A417** (1998) 391.
65. R. D. Cousins and V. L. Highland, Nuclear Instruments and Methods **A320** (1992) 331.
66. J. Heinrich, "Review of Banff challenge on upper limits", ref. [7], p. 125.
67. CDF Statistics Committee, "Recommendations concerning limits" (2005), http://www-cdf.fnal.gov/physics/statistics/recommendations/limits.txt.
68. A.Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione", G. Ist. Ital. Attuari, **4** (1933) 83.
N. Smirnov, "Table for estimating the goodness of fit of empirical distributions", Annals of Mathematical Statistics. **19** (2) (1948) 279, doi:10.1214/aoms/1177730256
69. T. W. Anderson and D. A. Darling, "Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes", Annals of Mathematical Statistics. **23** (1952) 193, doi:10.1214/aoms/1177729437; and "A Test of Goodness-of-Fit", Journal of the American Statistical Association, **49** (1954) 765, doi:10.2307/2281537
70. L. Lyons, D. Gibaut and P. Clifford, Nuclear Instr. Meth. **270** (1988) 210.
71. L. Lyons, A. Martin and D. Saxon, Phys Rev **D41** (1990) 982.
72. T. Trippe and Particle Data Group, private communication.
73. N. Suzuki et al, "Hubble Space Telescope cluster supernova study. V: Improving the Dark Energy constraints above $z > 1$ and building an early-type-hosted supernova sample",Astrophys J **746** (2012) 85, arXiv:1105.3470[astro-ph.CO]
74. L. Lyons and N. Wardle, "Statistical issues in searches for new phenomena in High Energy Physics", J Phys G Nucl Part Phys **48** (2018) 033001

75. ATLAS Collaboration, CMS Collaboration and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011", http://cds.cern.ch/record/1379837/files/NOTE2011_005.pdf
76. G. Cowan, K. Cranmer, E. Gross and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", Eur. Phys. J. C**71** (2011) 1554.
77. "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$", American Statistician **73** (2019)
78. L. Lyons, "Comparing two hypotheses" (1999), http://www-cdf.fnal.gov/physics/statistics/statistics_recommendations.html.
79. L. Lyons, "Methods for comparing two hypotheses", http://www.physics.ox.ac.uk/users/lyons/R_H_2009.pdf.
80. R. Trotta, Contemporary Physics **49** (2008) 71.
81. J. Heinrich, "A Bayes factor example: Poisson discovery", CDF note 9678 (2009), http://newton.hep.upenn.edu/~heinrich/bfexample.pdf.
82. S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses", Annals of Math. Stat. **9** (1938) 60.
83. R. Protassov et al., "Statistics: Handle with care. Detecting multiple model components with the likelihood ratio test", Astrophysics Journal **571** (2002) 545.
84. L. Demortier, "p-values and nuisance parameters", ref [7], p. 23.
85. L. Demortier, "Setting the scene for p-values" (2006), http://birs.pims.math.ca/~06w5054/Luc_Demortier.pdf.
86. S. G. Self and K. Y. Liang, JASA **82** (1987) 605.
87. M. Drton, "Likelihood ratio tests and singularities", Annals of Statistics **37** No. 2 (2009) 979, http://arxiv.org/abs/math/0703360.
88. K. Cranmer, "Statistics for LHC: progress, challenges and future", ref. [7], p. 47.
89. R. Cousins, J. Linnemann and J. Tucker, Nuclear Instr. Meth. A **595** (2008) 480.
90. E. Gross and O. Vitels, "Trial factors for the look elsewhere effect in high energy physics", E Phys J C **70** (2010) 525.
91. L. Lyons, "Discovering the Significance of 5σ" (2013), https://arxiv.org/abs/1310.1284
92. CDF Statistics Committee, "Frequently asked questions", http://www-cdf.fnal.gov/physics/statistics/statistics_faq.html#iptn4.
93. R. Cousins, "Annotated bibliography on some papers on combining significances or p-values", arXiv:0705.2209 (2007)
94. G. S. LaRue, J. D. Phillips and W. M. Fairbank, Phys. Rev. Lett. **46** (1981) 967.
95. J. R. Klein and A. Roodman, Annual Review of Nuclear and Particle Physics **55** (2005) 141.
96. I. Antcheva et al., "ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization", Computer Physics Communications, Anniversary Issue; **180** Issue 12 (2009) 2499:
W. Verkerke and D. Kirkby, "The RooFit toolkit for data modeling" (2003), arXiv:physics/0306116;
L. Moneta et al., "The RooStats Project", 13th Int. Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2010), arXiv:1009.1003.PoSACAT:057
97. V. Blobel, "Unfolding" in 'Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods' page 187 in [13].
98. M. Kuusela, "Uncertainty quantification in unfolding elementary particle spectra at the Large Hadron Collider", (2016) PhD thesis at EPFL Lausanne, https://infoscience.epfl.ch/record/220015/files/EPFL_TH7118.pdf

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 16
Integration of Detectors into a Large Experiment: Examples from ATLAS and CMS

Daniel Froidevaux

16.1 Introduction

16.1.1 The Context

The Large Hadron Collider (LHC) is the proton-proton accelerator which began operation in 2010 in the existing LEP tunnel at CERN in Geneva, Switzerland. It represents the next major step in the high-energy frontier beyond the Fermilab Tevatron (proton-antiproton collisions at a centre-of-mass energy of 2 TeV), with its design centre-of-mass energy of 14 TeV and luminosity of $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The high design luminosity is required because of the small cross-sections expected for many of the benchmark processes (Higgs-boson production and decay, new physics scenarios such as supersymmetry, extra dimensions, etc.) used to optimise the design of the general-purpose detectors over a period of 15 years or so. To achieve this luminosity and minimise the impact of simultaneous inelastic collisions occurring at the same time in the detectors (a phenomenon usually called pileup), the LHC beam crossings are 25 ns apart in time, resulting in 23 inelastic interactions per crossing on average at design luminosity. Two general-purpose experiments, ATLAS and CMS, were proposed for operation at the LHC in 1994 [1], and approved for construction in 1995. The experimental challenges undertaken by these two projects of unprecedented size and complexity in the field of high-energy physics, the construction and integration achievements realised over the years 2000–2008, and the expected performance of the commissioned detectors are described in a variety of detailed documents, such as the detector papers [2, 3]. In this chapter, much of the description of the lessons learned based on this huge effort, and of

D. Froidevaux (✉)
CERN, Geneva, Switzerland
e-mail: Daniel.Froidevaux@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_16

the comparisons in terms of expected performance have been taken and somewhat updated from a recent review [4]. For completeness, it is important to mention also the two more specialised and smaller experiments, ALICE [5] and LHCb [6]. In 2019, at a moment when the accelerator and experiments have just completed very successfully the so-called run-2 with 4 years of operation at a centre-of-mass energy of 13 TeV, and after run-1 with operation at lower energies topped with the discovery of the Higgs boson, it is interesting to look back not only on the period of construction and integration with its great expectations, a period which is the main focus of this chapter, but also on almost 10 years of operation and data-taking with its own challenges and of course with the excitement stemming from the analysis of real data.

The prime motivation of the LHC is to elucidate the nature of electroweak symmetry breaking, for which the Higgs mechanism is presumed to be responsible. The experimental study of the Higgs mechanism can also shed light on the consistency of the Standard Model at energy scales above 1 TeV. The Higgs boson is generally expected to have a mass below about 200 GeV [7]. This expectation could be relaxed if there are problems in the interpretation of the precision electroweak data [8] or if there are additional contributions to the electroweak observables [9]. A variety of models without Higgs bosons have also been proposed more recently, together with mechanisms of partial unitarity restoration in longitudinal vector boson scattering at the TeV scale [10]. All these possibilities may appear to be remote, but they serve as a reminder that the existence of a light Higgs boson cannot be taken for granted.

Theories or models beyond the Standard Model invoke additional symmetries (supersymmetry) or new forces or constituents (strongly-broken electroweak symmetry, technicolour). It is generally hoped that discoveries at the LHC could provide insight into a unified theory of all fundamental interactions, for example in the form of supersymmetry or of extra dimensions, the latter requiring modification of gravity at the TeV scale. There are therefore several compelling reasons for exploring the TeV scale and the search for supersymmetry is perhaps the most attractive one, particularly since preserving the naturalness of the electroweak mass scale requires supersymmetric particles with masses below about 1 TeV.

16.1.2 The Main Initial Physics Goals of ATLAS and CMS at the LHC

There have been many studies of the LHC discovery potential as a function of the integrated luminosity and the ones released just before data-taking [11, 12] have focussed on the first few years, over which about 10 fb^{-1} of integrated luminosity were expected to be accumulated by each experiment.

With some optimism that the performance of the ATLAS and CMS detectors would be understood rapidly and would be close to expectations, the expectations

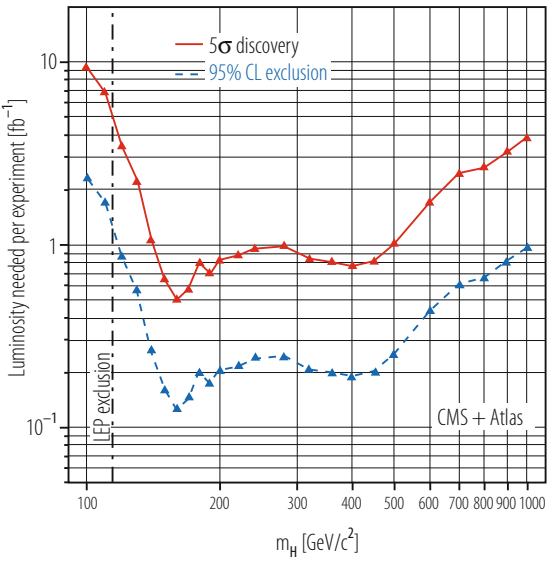


Fig. 16.1 Integrated luminosity required per experiment as a function of the mass of the Standard Model Higgs boson for a 5σ discovery or an exclusion at the 95% confidence level, combining the capabilities of ATLAS and CMS

at the time were that a Standard Model Higgs boson could be discovered at the LHC with a significance above 5σ over the full mass range of interest and for an integrated luminosity of only $5\,\text{fb}^{-1}$, as shown in Fig. 16.1. This discovery potential should, however, be taken with a grain of salt, since the evidence for a light Higgs boson of mass in the 110–130 GeV range would not only have to be combined over both experiments but also over several channels with very different final states ($H \rightarrow \gamma\gamma$ decays in association with various jet topologies, $tt\,H$ production with $H \rightarrow bb$ decay and $qq\,H$ production with $H \rightarrow \tau\tau$ decay). Achieving the required sensitivity in each of these channels would require an excellent understanding of the detailed performance of most elements of these complex detectors and would therefore require sufficient experimental data and time.

The discovery potential for supersymmetry was expected to be very substantial in the very first months of data-taking, since only $100\,\text{pb}^{-1}$ of integrated-luminosity would be sufficient to discover squarks or gluinos with masses below about 1.3 TeV [1, 11, 13], a large increase in sensitivity with respect to that ultimately achieved at the Tevatron. This sensitivity would increase to 1.7 TeV for an integrated luminosity of $1\,\text{fb}^{-1}$ and to about 2.2 TeV for $10\,\text{fb}^{-1}$, as shown in Fig. 16.2.

The few examples above illustrate the wide range of physics opened up by the seven-fold increase in energy from the Tevatron to the LHC. Needless to say, all Standard Model processes of interest, QCD jets, vector bosons and especially top quarks, would be produced in unprecedented abundance at the LHC, as illustrated in Table 16.1, and would therefore be studied with high precision by ATLAS and CMS.

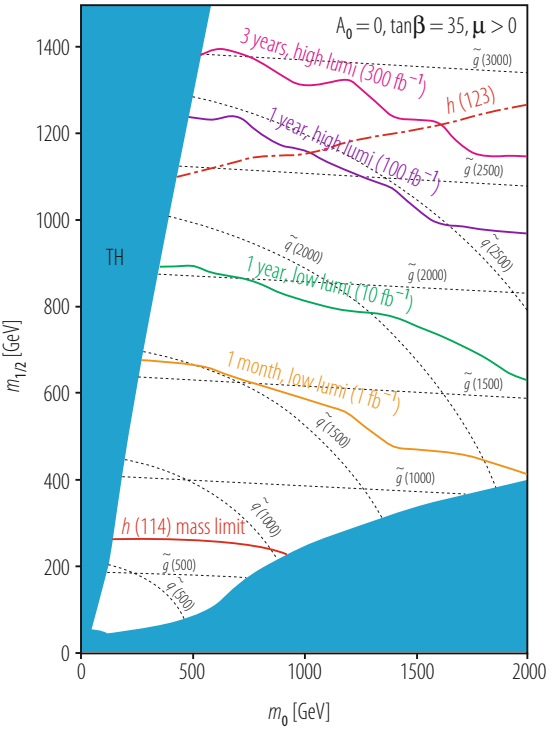


Fig. 16.2 Discovery potential for supersymmetry, expressed as lines corresponding to integrated luminosities ranging from 1 to 300 fb^{-1} in the (m_0, $m_{1/2}$) parameter plane, shown as an example for the CMS experiment. Also shown are lines representing constant squark or gluino masses. The discovery potential depends only weakly on the values assumed for tanβ, A_0 and the sign of μ

Table 16.1 For a variety of physics processes expected to be the most abundantly produced at the LHC, expected numbers of events recorded by ATLAS and CMS for an integrated luminosity of 1 fb^{-1} per experiment

Physics process	Number of events per 1 fb^{-1}
QCD jets with $E_T > 150$ GeV	10^6 (for 10% of trigger bandwidth)
$W \to \mu\nu$	$7.0 \cdot 10^6$
$Z \to \mu\mu$	$1.1 \cdot 10^6$
$tt \to e/\mu + X$	$1.6 \cdot 10^5$
Gluino-gluino production (mass about 1 TeV)	$10^2 \ldots 10^3$

16.1.3 A Snapshot of the Current Status of the ATLAS and CMS Experiments

From the year 2000 to end of 2009, the experiments have had to deal in parallel with a very complex set of tasks requiring a wide diversity of skills and personnel:

- the construction of the major components of the detectors was complete or nearing completion at the end of 2006, after a very long period of research

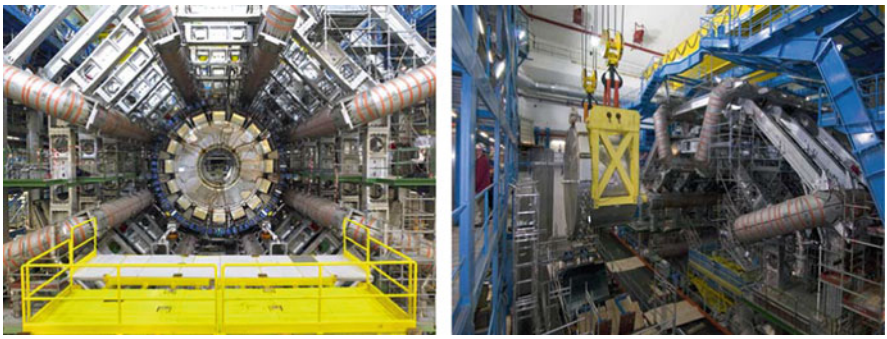


Fig. 16.3 Left: picture of the ATLAS barrel toroid superconducting magnet with its eight coils of 25 m length and of the ATLAS barrel calorimeter with its liquid Argon electromagnetic calorimeter and its scintillating tile hadronic calorimeter, as installed in the experimental cavern. Right: picture of the first end-cap LAr cryostat, including the electromagnetic, hadronic and forward calorimeters, as it is lowered into its docking position on one side of the ATLAS pit

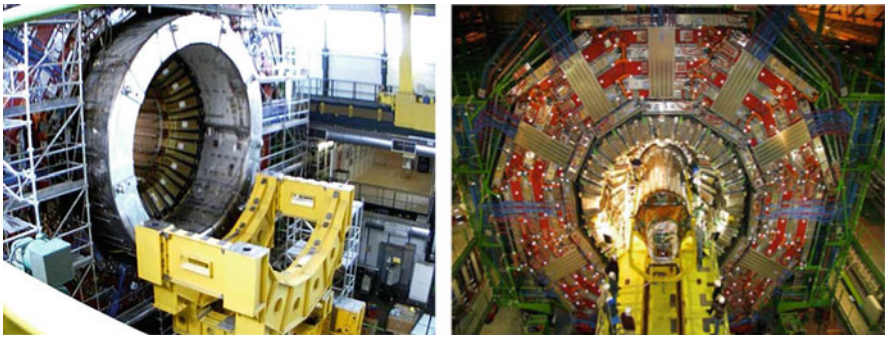


Fig. 16.4 Left: picture of the CMS superconducting solenoid, as integrated with the barrel muon system (outside) and with the barrel hadron calorimetry (inside). Right: picture of the insertion of the CMS silicon-strip tracker into the barrel crystal calorimeter

and development, including validation in terms of survival to irradiation and preparation of industrial manufacturing;

- the integration and installation phase began approximately in 2003 and extended all the way to 2007 for the last major components. ATLAS was being installed and commissioned directly in its underground cavern (see Fig. 16.3). In contrast, CMS is modular enough that it could be assembled above ground (see Fig. 16.4).
- the commissioning of the experiments with cosmic rays began in 2006, with the biggest campaigns in 2008 and 2009. These have yielded a wealth of initial results on the performance of the detectors in situ, a very important asset to ensure a rapid commissioning of the detectors for physics with collisions;
- the next commissioning step was achieved in an atmosphere of great excitement with first collisions at the injection energy of 900 GeV of the LHC machine and with very low luminosities of the order of $10^{26}-10^{27}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. All detector

components were able to record significant samples of data, albeit at low energy and with insufficient statistics to fully commission the trigger and reconstruction algorithms dedicated to provide the signatures required for the initial Standard Model measurements and searches for new physics.

In parallel with the rapidly evolving integration, installation and commissioning effort at the experimental sites, the collaborations have also reorganised themselves to evolve as smoothly and efficiently as possible from a distributed construction project with a strong technical co-ordination team to a running experiment with the emphasis shifting to monitoring of the detector and trigger operation, understanding of the detector performance in the real LHC environment and producing the first physics results. A small but significant part of the human and financial resources are already focusing on the necessary upgrades to the experiments required by the LHC luminosity upgrade programme.

This chapter has been structured in the following way: Sect. 16.2 presents an overview of the ATLAS and CMS projects in terms of their main design characteristics, describes briefly the magnet systems, and summarises the main lessons learned from the 15-year long research and development and construction period. The next three sections, Sects. 16.3–16.5, describe in more detail the main features and challenges related respectively to the inner tracker, to the calorimetry and to the muon spectrometer, in the specific case of the ATLAS experiment. The subsequent two sections, Sects. 16.6 and 16.7, discuss in broad terms the various aspects of, respectively, the trigger and data acquisition system and the computing and software, again in the context of the ATLAS experiment. The next section, Sect. 16.8, summarises and compares briefly the expected performances at the time of beginning of data-taking of the main ATLAS and CMS systems. The last and final section, Sect. 16.9, gives a very brief overview of the performance and physics results achieved over the past 10 years.

16.2 Overall Detector Concept and Magnet Systems

This section presents an overview of the ATLAS and CMS detectors, based on the main physics arguments which guided the conceptual design, and describes the magnet systems, which have driven many of the detailed design aspects of the experiments.

16.2.1 Overall Detector Concept

Figures 16.5 and 16.6 show the overall layouts respectively of the ATLAS and CMS detectors and Table 16.2 lists the main parameters of each experiment. Both experiments are designed somewhat as cylindrical onions consisting of:

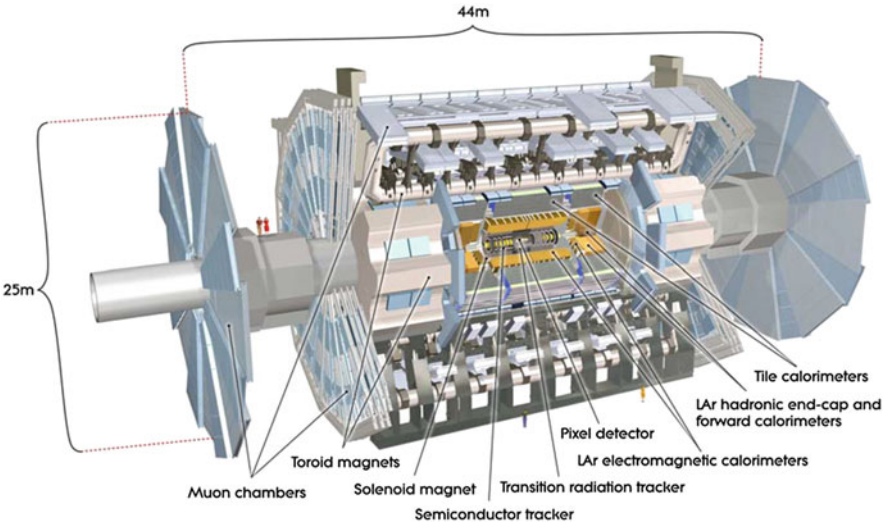


Fig. 16.5 Overall layout of the ATLAS detector

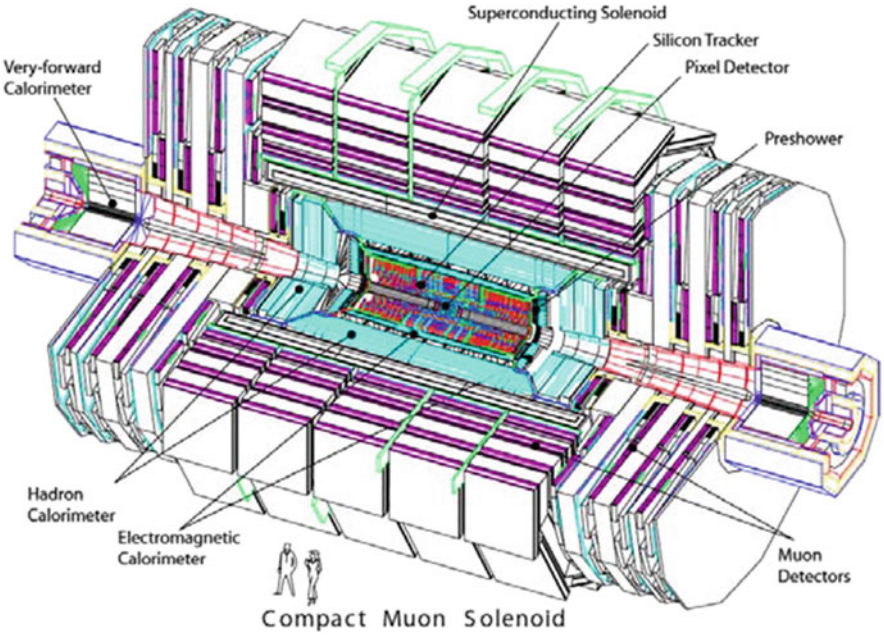


Fig. 16.6 Overall layout of the CMS detector

Table 16.2 Main design parameters of the ATLAS and CMS detectors

Parameter	ATLAS	CMS
Total weight [tons]	7000	12,500
Overall diameter [m]	22	15
Overall length [m]	46	20
Magnetic field for tracking [T]	2	4
Solid angle for precision measurements ($\Delta\phi \times \Delta\eta$)	$2\pi \times 5.0$	$2\pi \times 5.0$
Solid angle for energy measurements ($\Delta\phi \times \Delta\eta$)	$2\pi \times 9.6$	$2\pi \times 9.6$
Total cost (MCHF)	550	550

- an innermost layer devoted to the inner trackers, bathed in a solenoidal magnetic field and measuring the directions and momenta of all possible charged particles emerging from the interaction vertex;
- an intermediate layer consisting of electromagnetic and hadronic calorimeters absorbing and measuring the energies of electrons, photons and hadrons;
- an outer layer dedicated to the measurement of the directions and momenta of high-energy muons escaping from the calorimeters.

To complete the coverage of the central part of the experiments (often called barrel), so-called end-cap detectors (calorimetry and muon spectrometers) are added on each side of the barrel cylinders.

The sizes of ATLAS and CMS are determined mainly by the fact that they are designed to identify most of the very energetic particles emerging from the proton-proton collisions and to measure as efficiently and precisely as feasible their trajectories and momenta. The interesting particles are produced over a very wide range of energies (from a few hundred MeV to a few TeV) and over the full solid angle. They need therefore to be detected down to very small polar angles (θ) with respect to the incoming beams (a fraction of a degree, corresponding to pseudorapidities η of up to 5, where $\eta = -\log[\tan(\theta/2)]$; pseudorapidity is more commonly used at hadron colliders because the rates for most hard-scattering processes of interest are constant as a function of η). Most of the energy of the colliding protons is however dissipated in shielding and collimators close to the focussing quadrupoles (on each side of the experimental caverns, which house the experiments). The overall radiation levels will therefore be very high: many components in the detectors will become activated and will require special handling during maintenance, particularly near the beams.

For all the above reasons, both experiments have been designed following similar guiding principles:

- No particle of interest should escape unseen (except neutrinos, which will therefore be identified because their presence will cause an imbalance in the energy-momentum conservation laws governing the interactions measured in the experiments). The consequences of this simple statement are profound and far-

reaching when one goes beyond simple sketches and simulations to the details of the real experiment:

- successful operation of detectors able to measure the energies of particles with polar angles as small as one degree with respect to the incoming beams has required quite some inventiveness in material technology and a lot of detailed validation work to qualify the so-called forward calorimeters in terms of the very large radiation doses and particle densities encountered so close to the beams. Similar issues have been addressed of course very early on for the trackers, the main concerns being damage to semi-conductors (sensors and integrated circuits) and ageing of gaseous detectors. Even the muon detectors, to the initial surprise of the community, were confronted with irradiation and high-occupancy issues from neutron-induced cavern backgrounds pervading the whole experimental area;
- avoiding any cracks in the acceptance of the experiment (especially cracks pointing back to the interaction region) has been a challenge of its own in terms of minimising the thickness of the LAr cryostats in ATLAS and of properly routing the large number of cables required to operate the ATLAS and CMS inner trackers;
- if no particle can escape from the large volumes occupied by the experiments, then it becomes very hard for human beings to enter for rapid maintenance and repair. The access and maintenance scenarios for both experiments are quite complex and any major operation will only be feasible during long shutdowns of the accelerators. The detector design criteria have therefore become close to those required for space applications in terms of robustness and reliability of all the components.

- The high particle fluxes and harsh radiation conditions prevailing in the experimental areas have forced the collaborations to foresee redundancy and robustness for the measurements considered to be most critical. A few of the most prominent examples are described below:
  - CMS has chosen the highest possible magnetic field (4 T) combined with an inner tracker consisting solely of Silicon pixel detectors (nearest to the interaction vertex) and of Silicon microstrip detectors providing very high granularity at all radii. The occupancy of these detectors is below 2–3% even at the LHC design luminosity and the impact of pile-up is therefore minimal;
  - ATLAS has invested a very large fraction of its resources into three superconducting toroid magnets and a set of very precise muon chambers, constantly monitored with optical alignment devices, to measure the muon momenta very accurately over the widest possible coverage ($|\eta| < 2.7$) and momentum range (4 GeV to several TeV). This system provides a stand-alone muon momentum measurement of sufficient quality for all benchmark physics processes up to the highest luminosities envisaged for the LHC operation;

  - Both experiments rely on a versatile and multi-level trigger system to make sure the events of interest can be selected in real time at the highest possible efficiency.

- Efficient identification with excellent purity of the fundamental objects arising from the hard-scattering processes of interest is as important as the accuracy with which their four-momenta can be determined. Electrons and muons (and to a lesser extent photons and τ-leptons with their decay products) provide excellent tools to identify rare physics processes above the huge backgrounds from hadronic jets. The requirements at the LHC are far more difficult to meet than at the Fermilab Tevatron: for example, at a transverse momentum of 40 GeV, the electron to jet production ratio decreases from almost 10^{-3} at the Tevatron to a few 10^{-5} at the LHC, because of the much larger increase of the production cross section for QCD hadronic jets than for W and Z bosons.

  For reasons of size, cost and radiation hardness, both experiments have limited the coverage of their lepton identification and measurements to the approximate pseudorapidity range $|\eta| < 2.5$ (or a polar angle of 9.4° with respect to the beams). The implementation of these requirements has also had a very large impact on the design and technology choices of both experiments:

  - the length of the ATLAS and CMS super-conducting solenoids has been largely driven by the choices made for the lepton coverage;
  - ATLAS has chosen a variety of techniques to identify electrons, based first and foremost on the electromagnetic calorimeter with its fine segmentation along both the lateral and longitudinal directions of shower development, then on energy-momentum matching between the calorimeter energy measurement and the inner tracker momentum measurement, but enhanced significantly over most of the solid angle by the transition radiation tracker ability to separate electrons from charged pions. In contrast, CMS relies on the fine lateral granularity of its crystal calorimeter and on the energy-momentum matching with the inner tracker;
  - CMS has privileged the accuracy of the electron energy measurement with respect to the identification power with their choice of crystal calorimetry. The intrinsic resolution of the CMS electromagnetic (EM) calorimeter is superb with a stochastic term of 3–5.5% (see Sect. 16.8.2.1 for quantitative plots illustrating the performance) and the electron identification capabilities are sufficient to extract the most difficult benchmark processes from the background even at the LHC design luminosity.

- The overall trigger system of the experiments must provide a total event reduction of about 10^7 at the LHC design luminosity, since the number of inelastic proton-proton collisions will occur at a rate of about 10^9 Hz, whereas the storage capabilities will correspond to approximately 100 Hz for an average event size of 1–2 MBytes. Even today's state-of-the art technology is however far from approaching the performance required for taking a trigger decision in the very small amount of time between successive bunch crossings (25 ns).

The first level of trigger (or L1 trigger) in the ATLAS and CMS experiments is based on custom-built hardware extracting as quickly as possible the necessary information from the calorimeters and muon spectrometer and provides a decision in 2.5 to 3 μs, during which most of the time is spent in signal transmission from the detector (to make the trigger decision) and to the detector (to propagate this decision back to the front-end electronics). This reduces the event rate to about 100 kHz with a very high efficiency for most of the events of interest for physics analysis. During this very long (for relativistic particles) time, the hundreds of thousands of very sensitive and sophisticated radiation-hard electronics chips situated throughout the detectors have to store the successive waves of data produced every 25 ns in pipelines and keep track of the time stamps of all the data so that the correct information can be retrieved when the decision from the L1 trigger is received. The synchronisation of a vast number of front-end electronics channels over very large volumes has been a major challenge for the design of the overall trigger and timing control of the experiments.

16.2.2 Magnet Systems

The magnet systems of the ATLAS and CMS experiments [14] were at the heart of the conceptual design of the detector components and they have driven many of the fundamental geometrical parameters and of the broad technology choices for the components of the detectors. The large bending power required to measure muons of 1 TeV momentum with a precision of 10% has led both collaborations to choose superconducting technology for their magnets to limit the size of the experimental caverns and the overall costs. The choice of magnet system for CMS was based on the elegant idea of fulfilling at the same time with one magnet a high magnetic field in the tracker volume for all precision momentum measurements, including muons, and a high enough return flux in the iron outside the magnet to provide a muon trigger and a second muon momentum measurement for the experiment. This is achieved with a single solenoid of a large enough radius to contain most of the CMS calorimeter system. In contrast, the choice of magnet system for ATLAS was driven by the requirement to achieve a high-precision stand-alone momentum measurement of muons over as large an acceptance in momentum and η-coverage as possible. This is achieved using an arrangement of a small-radius thin-walled solenoid, integrated into the cryostat of the barrel electromagnetic calorimeter, surrounded by a system of three large air-core toroids, situated outside the ATLAS calorimeter systems and generating the magnetic field for the muon spectrometer. The main parameters of these magnet systems are listed in Table 16.3 and their stored energies are compared to those of previous large-scale magnets in high-energy physics experiments in Fig. 16.7.

In CMS, the length of the solenoid was driven by the need to achieve excellent momentum resolution over the required η-coverage and its diameter was chosen such that most of the calorimetry is contained inside the coil. In ATLAS, the

Table 16.3 Main parameters of the CMS and ATLAS magnet systems

Parameter	CMS solenoid	ATLAS Solenoid	ATLAS Barrel toroid	ATLAS End-cap toroids
Inner diameter	5.9 m	2.4 m	9.4 m	1.7 m
Outer diameter	6.5 m	2.6 m	20.1 m	10.7 m
Axial length	12.9 m	5.3 m	25.3 m	5.0 m
Number of coils	1	1	8	8
Number of turns per coil	2168	1173	120	116
Conductor size [mm^2]	64 × 22	30 × 4.25	57 × 12	41 × 12
Bending power	4 T · m	2 T · m	3 T · m	6 T · m
Current	19.5 kA	7.6 kA	20.5 kA	20.0 kA
Stored energy	2700 MJ	38 MJ	1080 MJ	206 MJ

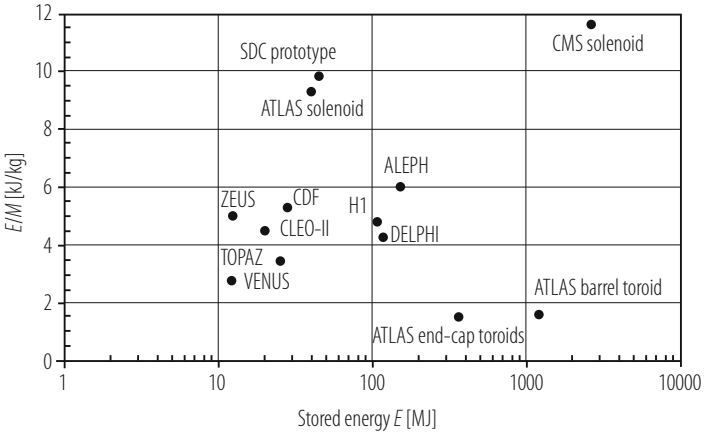


Fig. 16.7 Ratio of stored energy over mass, E/M, versus stored energy, E, for various magnets built for large high-energy physics experiments

position of the solenoid in front of the barrel electromagnetic calorimeter has demanded a careful optimisation of the material in order to minimise its impact on the calorimeter performance and its length has been defined by the design of the overall calorimeter and inner tracker systems, leading to significant non-uniformity of the field at the end of the tracker volume.

The main advantages and drawbacks of the chosen magnet systems can be summarised as follows, considering successively the inner tracker, calorimeter and muon system performances (see Sect. 16.8):

- the higher field strength and uniformity of the CMS solenoid provide better momentum resolution and better uniformity over the full η-coverage for the inner tracker;

- the position of the ATLAS solenoid just in front of the barrel electromagnetic calorimeter limits to some extent the energy resolution in the region $1.2 < |\eta| < 1.5$;
- the position of the CMS solenoid outside the calorimeter limits the number of interaction lengths available to absorb hadronic showers in the region $|\eta| < 1$;
- the muon spectrometer system in ATLAS provides an independent and high-accuracy measurement of muons over the full η-coverage required by the physics. This requires however an alignment system with specifications an order of magnitude more stringent (few tens of μm) than those of the CMS muon spectrometer. In addition, the magnetic field in the ATLAS muon spectrometer must be known to an accuracy of a few tens of Gauss over a volume of close to 20,000 m^3. The software implications of these requirements are non-trivial (size of map in memory, access time);
- the muon spectrometer system in CMS has limited stand-alone measurement capabilities and this affects the triggering capabilities for the luminosities envisaged for the LHC upgrade.

In terms of construction, the magnet systems have each turned out to be a major project in its own right with very direct and strong involvement from the Technical Coordination team [15] and from major national laboratories and funding agencies. A detailed account of the construction of these magnets is beyond the scope of this review and this section can be concluded by simply stating that during the course of the past few years, all these magnets have undergone very successfully extensive commissioning steps, sustained operation at full current, in particular for cosmic-ray data-taking in 2008/2009, and stable operation with beam in the LHC machine at the end of 2009.

16.2.2.1 Radiation Levels

At the LHC, the primary source of radiation at full luminosity comes from collisions at the interaction point. In the tracker, charged hadron secondaries from inelastic proton-proton interactions dominate the radiation backgrounds at small radii while further out other sources, such as neutrons, become more important. Table 16.4 shows projected radiation levels in key areas of the detector.

In ATLAS, most of the energy from primaries is dumped into two regions: the TAS (Target Absorber Secondaries) collimators protecting LHC quadrupoles and the forward calorimeters. The beam vacuum system spans the length of the detector and in the forward region is a major source of radiation backgrounds. Primary particles from the interaction point strike the beam-pipe at very shallow angles, such that the projected material depth is large. Studies have shown that the beam-line material contributes more than half of the radiation backgrounds in the muon system. The deleterious effects of background radiation fall into a number of general categories: increased background and occupancies, radiation damage and ageing of

Table 16.4 The 1 MeV neutron equivalent fluence (F_{neq}) and doses in key areas of the ATLAS detector after 500 fb^{-1} of data (estimated to be approximately 7 years of operation at design luminosity)

Inner detector					
Location	F_{neq}		Dose		Charged-particle flux
	[10^{14} cm^{-2}]		[kGy]		above 10 MeV [Hz/cm^2]
Pixel layer 0	13.5		790		$40 \cdot 10^6$
SCT layer 1	0.8		38		$1.5 \cdot 10^6$
SCT disk 9	0.6		23		10^6
TRT outer radius	0.25		3.5		10^5
Calorimeters					
Location	$\lvert\eta\rvert$		Maximum dose [kGy]		
EM barrel	1.475		1.2		
EM end-cap	3.2		150		
Tile	1.2		0.15		
HEC	3.2		30		
FCal	4.9		1000		
Muon spectrometer					
Location	Flux				Single-plane rates
	[kHz/cm^2]		[Hz/cm^2]		[Hz/cm^2]
	n	γ	μ	p	
Barrel chambers	2.6–4.0	1.0–1.5	0.3–4.5	0.4–3.2	6.0–11.0
Inner edge of inner wheel	79	25	21	64	347
Inner edge of outer wheel	2.7	1.5	3	0.9	12

Also given are the charged-particle fluxes in the tracker and fluxes and single-plane rates in the muon spectrometer

detector components and electronics, single-event upsets and single-event damage, and creation of radionuclides which will impact access and maintenance scenarios.

16.2.3 Lessons Learned from the Construction Experience

It is fair to say that most of the physicists and engineers involved in the ATLAS and CMS construction were faced with a challenge of this scope and size for the first time. It seems therefore appropriate to put some emphasis in this article on the lessons learned from the construction of these detectors. This section describes the general lessons learned and the next sections will give more explicit examples in many cases when describing the experience from the construction of the detector components.

The lessons learned are of varying nature, many are organisational, many are technical and some are sociological. Some are specific to the LHC, some are specific to the way international high-energy physics collaborations work, and some are of a

general enough nature that they might well apply to any complex high-tech project of this size. It is therefore hard to classify them in a clear logical order, and this review has attempted to rank them from the general and common to the specific and unique to the LHC.

16.2.3.1 Time-Scales, Project Phases and Schedule Delays

If there has been one lesson learned from the days in the early 1990s when ATLAS and CMS came into being as detector concepts, it is certainly that the research and development phase of projects of this complexity are impossible to plan with real certainty about the time-scales involved. Modern tools for project management are of little help here because the vagaries of the initial phase do not generally obey the simple laws of project schedules and charts. These can be a posteriori explained of course:

- the research and development phase for new high-tech detector elements, such as radiation-hard silicon sensors and micro-electronics, crystals grown from a new material, large-scale electrodes for operation at high voltage in liquid Argon, etc., will always be a phase to which one has to allocate as much time as feasible within the overall project schedule constraints. The justification for this is basically that the potential rewards are enormous, as was exemplified by the late but striking success of the deep sub-micron micro-electronics chips pioneered by CMS and now used throughout all LHC experiments, and by the late but successful operation of CMS $PbWO_4$ crystals with their avalanche photodiode readout and associated electronics. Making the appropriate research and development choices at the right time will however always remain a challenge for any new project of this scope and complexity.
- less known to many colleagues in our community is the phase during which the components for producing complex detector modules are launched for manufacturing in industry. This phase can indeed be planned correctly if the required physicist/engineering experience is available, if the funding allows for multiple suppliers to mitigate potential risks, and if the physicists agree quickly to moderate their usually very demanding specifications to adapt them to the actual capabilities of industry.

  Experience has shown however that success was far from guaranteed in this phase, with causes for delays or outright initial failures ranging from being forced to award contracts to the lowest bidder, to incomplete technical specifications, to handling and packaging issues during manufacturing, particularly for polyimide-based products, of which there are many thousands of m^2 in both experiments. This material shows up under various forms (especially in flexible printed circuit boards for various applications) and is a basic insulating material with excellent electrical and mechanical properties, with very high tolerance to radiation, but unfortunately also with a high propensity to absorb moisture and thereby lead to unexpected changes in even the course of a well-defined manufacturing process.

Serious technical problems in this area have affected the manufacturing schedule of major components of both experiments (hybrids for semi-conductor detectors, flexible parts of printed-circuit boards, large-size electrodes for electromagnetic calorimetry), but other issues such as welding, brazing and general integrity and leak-tightness of thin-walled cooling pipes have also been a concern for several of the components in each experiment.

In addition, several of the more significant contracts were seriously affected by changes in the industrial boundary conditions (insolvency, change of ownership). The recommended purchasing strategy of having multiple suppliers for large contracts, to minimise the consequences from a possible failure in the case of a single supplier, has not always been the optimal one (high-quality silicon sensors are perhaps the most prominent example).

The detailed construction planning can be consulted in the various Technical Design Reports (TDR), most of which were submitted from 1996 to 1998 to seek approval for construction of the major detector components. This called for completion of this construction phase by mid-2001 to mid-2003. At the time when a big schedule and financial crisis shook the LHC project in fall 2001 (see below), it was already clear that many detector components would not be on schedule by a significant margin.

The 2-year delay in the completion of the accelerator resulting from this crisis was also needed by the experiments, as can be seen from Table 16.5, which illustrates the major construction milestones originally planned at the time of the TDRs and actually achieved. When trying to assess the significance of the differences between the dates achieved for the delivery of major components of

Table 16.5 Main construction milestones for the ATLAS and CMS detectors

Detector system	ATLAS		CMS	
	TDR	Actual	TDR	Actual
Pixels	06/03	03/07	03/05	12/07
Silicon micro-strips (barrel)	12/02	07/05	03/04	10/06
Silicon micro-strips (end-caps)	12/02	06/06	03/04	10/06
Transition radiation tracker	03/04	12/05		
Electromagnetic calorimeter (barrel)	06/03	07/04	12/03	03/07
Electromagnetic calorimeter (end-caps)	01/04	09/05	06/04	03/08
Hadronic calorimeter	12/02	02/04	12/03	12/04
Muon chambers	12/04	12/05	12/03	06/06
Solenoid magnet	01/02	09/01	03/03	12/05
Barrel toroid magnet	06/02	06/05		
End-cap toroid magnet	12/03	11/06		

Shown are the milestones for the delivery of major components to CERN, as planned at the time of the Technical Design Reports (TDR), and the actual delivery milestones achieved

the experiments and those planned 9 years ago, it is important to remember the prominent events, at CERN and within the collaborations, which happened during these years:

- at the time of the submission of the various TDRs for ATLAS and CMS, the construction and installation schedule was worked out top-down, based on a ready-for-operation date of summer 2005 for the LHC machine and the experiments;
- in 1999, the CMS collaboration decided to replace the micro-strip gas chamber baseline technology for the outer part of their Inner Detector by "low-cost" silicon micro-strip detectors. This is probably the most outstanding example of decisions, which the collaborations had to take after the TDRs were submitted and which have affected the construction schedule in a major way;
- in 2001, when the CERN laboratory management announced significant cost overruns, mostly in the machine, but also in the ATLAS and CMS experiments, it also announced a 2-year delay in the schedule for the machine, which obviously led to a readjustment of the construction and installation schedule of the experiments. By that time, both in ATLAS and CMS, the Technical Co-ordination teams had worked out a realistic installation schedule, which still needed to be fleshed out substantially in areas such as services installation, commissioning of ancillary equipment for operation of the huge devices to be operated underground, etc.;
- the ATLAS experimental cavern was delivered more or less on time in spring 2003, whereas the CMS experimental cavern suffered considerable delays and was delivered only towards the end of 2004.

16.2.3.2 Physicists and Engineers: How to Strike the Right Balance?

This is a very delicate issue because there exists no precise recipe to solve this problem. The ATLAS and CMS experiments were born from the dreams of physicists but are based today on the calculations and design efforts from some of the best teams of engineers and designers in the world. One should not forget that, originally (in 1987), even the physicists thought that only a muon spectrometer behind an iron dump was guaranteed to survive the irradiation and that most tracking technologies were doomed at the highest luminosities of the LHC [16].

Although a strong central and across-board (from mechanics to electronics, controls and computing) engineering effort would have been desirable from the very start (i.e. around 1993), a standard centralised and very systematic engineering approach alone, as is frequently used in large-scale astronomy projects, could not have been used for several reasons:

- the cost would have been prohibitive;
- only the physicists can actually make the sometimes difficult choices and decisions when faced with problems requiring certain heart-wrenching changes in the fundamental parameters of the experiment (number of layers in the

tracking detectors, number of cells in the electromagnetic calorimeter, overall strength and uniformity of the magnetic field, etc.). The number of coils to be constructed in the ATLAS superconducting toroid and the peak field of the CMS central solenoid are two examples of early and fundamental parameters of the experiments, which were studied for quite some time and had a significant bearing on the overall cost of the experiments;

- some of the usual benefits of such an approach, such as optimised production costs for repetitive manufacturing of the same product, are not there to be reaped when considering the experiments as a whole rather than looking at individual components, such as the micro-strip silicon modules, which number in many thousands and did indeed benefit in many aspects from a systematic engineering approach;
- the overall technological scope of these nascent experiments required creativity and novel approaches in areas as far apart as 3D-calculations of magnetic fields and forces over very large volumes containing sometimes unspecified amounts of magnetic materials and radiation-dose and neutron-fluence calculations of unprecedented complexity in our field to evaluate the survival of a variety of objects, from the basic materials themselves to complex micro-electronics circuits. Only a well-balanced mix of talented and dedicated designers, engineers and physicists could have tackled such issues with any chance of success;
- the decision-making processes in our community cannot be too abrupt. Consensus needs to be built, especially between physicists but also between engineers from sometimes widely different cultures and backgrounds.

In retrospect, however, there has emerged as a clear lesson, that the management of the experiments should have evolved at an earlier stage the decision-making process from a physicist-centric one at the beginning, when little was known about the detailed design of all the components, to a more engineer-centric one, as the details were fleshed out more and more. Establishing engineering envelopes and assembly drawings for the different systems, routing the very large and diverse amount of services needed to operate complex detectors distributed everywhere across the available space, and designing, validating and procuring common solutions for many of the electronics and controls components are examples, which clearly illustrate this need. The collaborations have indeed encountered difficulties to recognise such needs and to react to them at the appropriate moment in time.

16.2.3.3 International and Distributed: A Strength or a Weakness?

ATLAS and CMS are truly international and distributed collaborations, even if the engineering and/or manufacturing of some of the major components of both experiments have been entrusted to large laboratories situated all across the world. Modern technology (web access to document servers, video-conferencing facilities, more uniform standards, such as the use of the metric system, for drawings, specifications and quality assurance methods, electronic reporting tools) has been instrumental in

improving the efficiency of the various strands of these collaborations, an admittedly weak point of such organisations. There are two major weaknesses intrinsic to collaborations structured as ATLAS and CMS with distributed funding resources:

- one is that it is not simple to converge on the minimum required number of technologies once the research and development phase is over. One example of perhaps unnecessary multiplication of technologies are the precision chambers in the ATLAS muon spectrometer, where the highest-η part of the measurements are covered by cathode strip chambers rather than the monitored drift tube technology used everywhere else. A similar example can be found in the CMS muon spectrometer, which is also equipped with two different chamber technologies in the barrel and end-cap regions (see Sect. 16.5).
- the decision-making process is sometimes skewed by the difficulty of conveying a global vision of the best interests of the project, which should be weighed against the more localised and focussed interests of particular funding agencies, some of which operate within a rather inflexible legal framework.

The strengths of this international and distributed approach far outweigh however its deficiencies over a much more centralised one, such as that adopted for the Super-Conducting Super Collider with a centralised funding and management in Waxahachie (Texas) about 15 years ago:

- the flexibility achieved has often provided solutions to the inevitable problems, which have shown up during the design and construction phase. Whenever a link in the chain was shown to falter or even to be totally missing, the collaboration has often been able to find alternate solutions. If a large laboratory had difficulties in meeting a complex technological challenge alone because of limitations in funding and human resources, other laboratories with similar expertise could be sought out and integrated into the effort with minimal disruption. If the production line for certain detectors did not churn out the required number of modules per unit time because of yield issues or of an underestimate of the human resources required, other production lines, often on different continents with cheaper labour costs, were launched and operated successfully.
- many concrete examples have shown that motivation and dedication to the project go together with the corresponding responsibilities, both technical and managerial. It is worthwhile also to note here that it surely would have been beneficial for the overall LHC project if the management of the ATLAS and CMS experiments would have been integrated as a real partner into the CERN management structure at the highest level right from the beginning. Both experiments were severely handicapped by a cost ceiling without contingency defined top-down more than 10 years ago.

  It is fair to say that, without the motivation and dedication of many of our colleagues all over the world, who fought and won their own battles at all required levels (technical, funding, human resources, organisational), and of their funding agencies, the construction of ATLAS and CMS would not have reached its astounding and successful completion with only small parts of each experiment

deferred. Dealing with significant deferrals has always been damaging to the atmosphere of large collaborations of this type and the fact that both experiments are now essentially complete should certainly be attributed to the credit of all their participants.

A particular mention should go here to our Russian colleagues, who have not only strongly contributed intellectually to the experiments, as all the others, from the very beginning, but who also staffed continuously, together with other Eastern European colleagues and also colleagues from Asia, a very large fraction of the teams needed to assemble, equip, test and commission the major detector components. This was quite striking during the installation period from just listening to the conversations occurring in the lifts bringing people and equipment up and down the experimental shafts.

- the concept of deliverables has also turned out to the advantage of the projects. Each set of institutes in each country have been asked to deliver a certain fraction of specific components of the detector systems, ranging from a modest (but critical!) scope, such as the fabrication of the C-fibre cylinders for the barrel semi-conductor tracker in ATLAS, to a very large (and very visible to the whole collaboration!) scope, such as the CMS crystal production in several commercial companies, or as the ATLAS super-conducting solenoid built in Japanese industry, in close collaboration with institutes from the same country, which are full-fledged members of the collaboration.

  This concept has certainly maximised the overall funding received by ATLAS and CMS, because each funding agency has to a certain extent been asked and has agreed to take responsibility for the delivery of certain detector components without assigning to these a specific cost, since the real costs vary from country to country, and even the ratios of costs between different countries inevitably vary, because of the approximately uniform costs of raw materials as compared to the wildly differing costs of skilled and unskilled labour. Since the infrastructure of the experiments is a mixture of low and high technology components, most participating countries have in the end been able to contribute efficiently in kind to the common projects of interest to the whole collaboration.
- the scheme based on deliverables rather than raw funding could not have worked however without being completed by a sizable set of common projects, to which the funding agencies had to contribute, either through funds to be handled by the management of the experiments, either through in-kind contributions, the cost of which was determined in the context of the same scheme as for the deliverables. Examples of these common projects are the magnets of both experiments, the LAr cryostats and cryogenics of ATLAS, and much of the less high-tech infrastructure components of both experiments.
- finally, the computing operations of the experiments and the analysis of the data taken over the next 10 years do and will require a very distributed and international style of working also. This is not really new to our community, it is just of an unprecedented scale in size and duration. The collaborations are evolving now from an organisational model focussed initially on research and development and then on construction to a new model, which is focussed more on

detector operation, monitoring of the data quality and data preparation, leading to the analysis work required to understand precisely the behaviour of the detectors and extract as efficiently as possible the exciting physics ahead of us. The years spent together and the difficulties overcome over a 15-year long period of design and construction have certainly cemented the collaborations in a spirit of respect and mutual understanding of all their diverse components. This will surely turn out to be an excellent preparation for the forthcoming challenges when faced with real experimental data.

16.2.3.4 A Well Integrated and Strong Technical Co-ordination Team

It is clear that without such a team the experiments would most probably have faced insurmountable construction delays and integration problems. The Technical Co-ordination team must in a sense be perceived as the strong backbone of the experiment by all the physicists in the community. This was indeed the case in the installation phase of the experiments, at a time when it had to smoothly execute a complex suite of integration and installation operations for detector components arriving from all over the world. But this was less the case 10–15 years ago, at a time when the physicists and engineers in this team were sometimes perceived as a nuisance disrupting the delicate balance of the collaboration and were criticised in different ways:

- many physicists and engineers had great trouble when asked to specify all the details of cables, pipes and connectors, at a very early time (15 years ago) when they were desperately trying to move into mass production;
- strong resistance to reviews was encountered, based on partially correct, but also partially fallacious, arguments that all the expertise in a given area was already available in the project under review;
- the multiplicity of reviews also caused sometimes considerable friction and frustration, especially since an overall co-ordination between funding agency reviews and internal project reviews was almost impossible to put into place.

In retrospect, these reviews are indeed necessary, whether or not all of their recommendations and outcomes have turned out to be of a specific concrete usefulness, because they have usually forced the project teams to collect documentation, take stock, step back and think about issues sometimes obscured by the more immediate and pressing problems at hand.

Although the construction of the individual detector components can be argued to have been quite successful under the umbrella of deliverables and in the absence of a fully centralised management of the experiment resources, there are obviously a variety of tasks, which have to be solved by a strong centralised team of designers, engineers and physicists. As in any such process, this team is much better accepted if it is built up at least partially from people within the collaboration, who are already well integrated in and known to the collaboration. Despite all the grumbling and

moaning, the efforts of the Technical Co-ordination team have been crucial to the success of the ATLAS and CMS projects:

- finding common (often commercial) solutions does not come easily to large numbers of inventive and often opinionated physicists. Common solutions across the experiments are even harder to achieve, although they have turned out to be profitable to all parties in a number of areas. Clearly the strong research and development programme launched in 1989 by CERN for the development of the LHC detector technologies has been a key element in the definition of the various detector concepts (radiation-hard silicon detectors and electronics, electromagnetic and hadronic calorimetry, various tracking technologies, etc.).

  In the areas where such common (often commercial) solutions have been adopted in many cases in the past, the successes of the research and development programme have been less spectacular (data transmission, specialised trigger processors, various offline software developments), most probably because the solutions emerging today were not easy to predict from the technology trends of 20 years ago, when the worldwide web, mobile phones, inexpensive desktop computing and high-speed networks did not exist.

  The Technical Co-ordination team has certainly been very instrumental in encouraging the collaboration to adopt common technical solutions and has also delegated to the appropriate persons in the collaboration the mandate to negotiate and agree these common solutions across the experiments: the frame contracts with major micro-electronics suppliers, the gas systems, the power supplies, the electronics crates and racks and the slow controls infrastructure hardware and database software can be quoted as some of the more prominent examples.
- establishing a strong quality assurance and review process across the whole collaboration is a must at an early stage in such complex projects, where standard commercial products have often failed, sometimes for multiple reasons owing to the boundary conditions in the experimental caverns (radiation background and magnetic field).

  As stated above, the review process (from conceptual engineering design reviews, to production readiness and production advancement reviews) can be very beneficial and even well accepted within the collaboration if it is kept lightweight and perceived as executed by people involved in the project as all the others rather than by an elite breed of top-level managers.

  Most of the ATLAS and CMS Technical Design Reports quoted as references in this review address quality assurance with ambitions and specifications, which are fully justified on paper but much harder to implement in reality when facing time pressure and the inevitable lack of human resources to fulfill every aspect of the task. In relation to industry in particular, the effort required in monitoring production of delicate components had been totally underestimated or even ignored in the design phase. The reviews put in place by the Technical Co-ordination team have played an important role in keeping all aspects related to schedule, resources and quality assurance under control during the detector construction. They have also ensured that large groups with significant project

responsibilities were not allowed to operate for too long in a stand-alone mode without synchronising with and reporting back to Technical Co-ordination, the management of the experiments and the collaboration at large. The risks involved in letting things go astray too much are simply unacceptable for projects of this complexity and size.

- As stated above, one weakness perhaps of the multiple dimensions under which ATLAS and CMS are viewed is that the funding agencies have often conducted their own necessary review processes in a way largely decoupled from the review process operated by the management of the experiments. This weakness stems from the lack of central control of expenditures because of the distributed funding and spending responsibilities. This can obviously lead to inefficiencies in the actual execution of the project and, worse, sometimes to conflicting messages given to the institutes concerning priorities, since those of a given funding agency may not always coincide with those of the experiment. The common funds necessary to the construction of significant components of the experiments, such as magnets, infrastructure, shielding, cryostats, etc., are a prominent example which comes to mind, when assessing which of the components of the experiments had the most difficulty in dealing with the multi-threaded environment, in which the detector construction has been achieved.

Finally, it is in the very recent phase of assembly, installation and commissioning of the ATLAS and CMS detectors that the enormous efforts and contribution from the Technical Co-ordination teams have been most visible: they have had to organise the vast teams of sub-contractors and specialised personnel from the collaborating institutes and they have had to deal with the daily burden of making sure all the tasks were executed as smoothly as possible with safety as one of the paramount requirements.

16.3 Inner Tracking System

16.3.1 Introduction

The ATLAS tracker is designed to provide hermetic and robust pattern recognition, excellent momentum resolution and both primary and secondary vertex measurements [17] for charged tracks above a given p_T threshold (nominally 0.5 GeV, but as low as 0.1 GeV in some ongoing studies of initial measurements with minimum-bias events) and within the pseudorapidity range $|\eta| < 2.5$. It also provides electron identification over $|\eta| < 2.0$ and a wide range of energies (between 0.5 and 150 GeV). It is contained within a cylindrical envelope of length ± 3512 mm and of radius 1150 mm, within the solenoidal magnetic field of 2 T. Figures 16.8 and 16.9 show the sensors and structural elements traversed by 10 GeV tracks in respectively the barrel and end-cap regions.

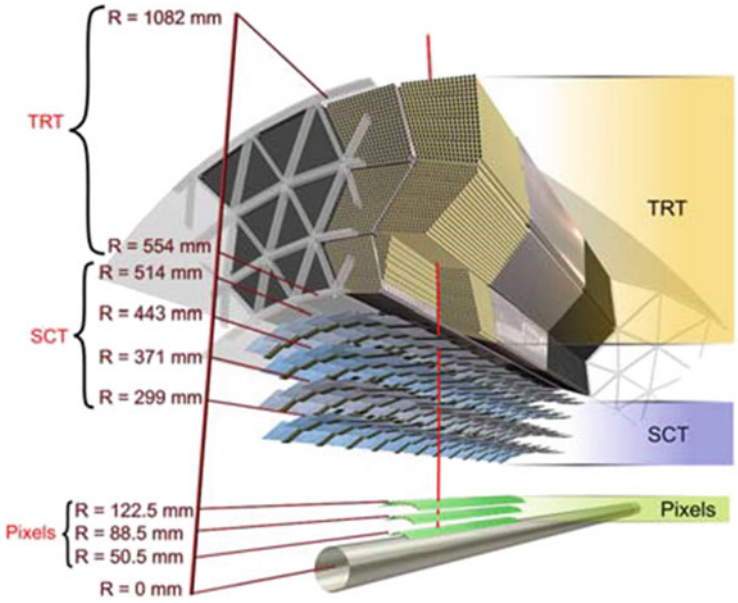


Fig. 16.8 Drawing showing the sensors and structural elements traversed by a charged track of 10 GeV p_T in the ATLAS barrel inner detector ($\eta = 0.3$). The track traverses successively the beryllium beam-pipe, the three cylindrical silicon-pixel layers with individual sensor elements of $50 \times 400\,\mu m^2$, the four cylindrical double layers (one axial and one with a stereo angle of 40 mrad) of barrel silicon-microstrip sensors (SCT) of pitch 80 μm, and approximately 36 axial straws of 4 mm diameter contained in the barrel transition-radiation tracker modules within their support structure

The ATLAS tracker consists of three independent but complementary sub-detectors. At inner radii, high-resolution pattern recognition capabilities are available using discrete space-points from silicon pixel layers and stereo pairs of silicon micro-strip (SCT) layers. At larger radii, the transition radiation tracker (TRT) comprises many layers of gaseous straw tube elements interleaved with transition radiation material. With an average of 36 hits per track, it provides continuous tracking to enhance the pattern recognition and improve the momentum resolution over $|\eta| < 2.0$ and electron identification complementary to that of the calorimeter over a wide range of energies.

Table 16.6 lists the main parameters of the ATLAS tracker:

- the radial position of the innermost measurement is essentially determined by the outer diameter of the beam pipe, which has been manufactured using expensive and delicate Beryllium material over an overall length of 7 m. The active part of the tracker has a half-length of 280 cm, slightly longer than that of its solenoid, resulting in significant field non-uniformities and momentum resolution degradation at each end.

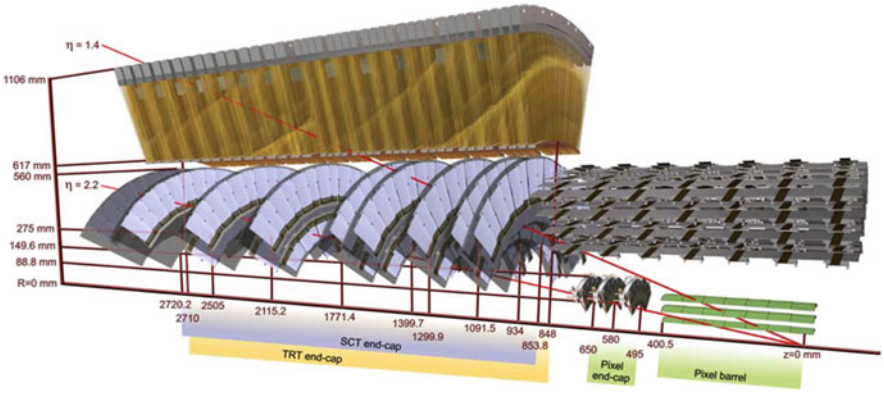


Fig. 16.9 Drawing showing the sensors and structural elements traversed by two charged tracks of 10 GeV p_T in the ATLAS end-cap inner detector ($\eta = 1.4$ and 2.2). The end-cap track at $\eta = 1.4$ traverses successively the beryllium beam-pipe, the three cylindrical silicon-pixel layers with individual sensor elements of $50 \times 400\,\mu m^2$, four of the disks with double layers (one radial and one with a stereo angle of 40 mrad) of end-cap silicon-microstrip sensors (SCT) of pitch ~80 μm, and approximately 40 straws of 4 mm diameter contained in the end-cap transition radiation tracker wheels. In contrast, the end-cap track at $\eta = 2.2$ traverses successively the beryllium beam-pipe, only the first of the cylindrical silicon-pixel layers, two end-cap pixel disks and the last four disks of the end-cap SCT. The coverage of the end-cap TRT does not extend beyond $|\eta| = 2$

Table 16.6 Main parameters of the ATLAS tracker system

Item		Radial extension [mm]	Length [mm]
Overall tracker envelope		$0 < R < 1150$	$0 < \|z\| < 3512$
Beam-pipe		$29 < R < 36$	
Pixel	Overall envelope	$45.5 < R < 242$	$0 < \|z\| < 3092$
3 cylindrical layers	Sensitive barrel	$50.5 < R < 122.5$	$0 < \|z\| < 400.5$
2 × 3 disks	Sensitive end-cap	$88.8 < R < 149.6$	$495 < \|z\| < 650$
SCT	Overall envelope	$255 < R < 549$ (barrel)	$0 < \|z\| < 805$
		$251 < R < 610$ (end-cap)	$810 < \|z\| < 2797$
4 cylindrical layers	Sensitive barrel	$299 < R < 514$	$0 < \|z\| < 749$
2 × 9 disks	Sensitive end-cap	$275 < R < 560$	$839 < \|z\| < 2735$
TRT	Overall envelope	$554 < R < 1082$ (barrel)	$0 < \|z\| < 780$
		$617 < R < 1106$ (end-cap)	$827 < \|z\| < 2744$
73 straw planes	Sensitive barrel	$563 < R < 1066$	$0 < \|z\| < 712$
160 straw planes	Sensitive end-cap	$644 < R < 1004$	$848 < \|z\| < 2710$

- the total power required for the tracker front-end electronics will increase from approximately 62 to 85 kW from initial operation to high-luminosity operation after irradiation. Bringing this amount of power to the detector requires large amounts of copper; the resulting heat load is very uniformly distributed across the entire active volume of the tracker and has to be removed using innovative techniques (fluor-inert liquids to mitigate the risks from possible leaks, thin-walled pipes made from light metals, evaporative techniques for optimal heat removal in the case of the silicon-strip and pixel detectors). There is also considerable heat created by the detectors themselves: the silicon-strip modules will dissipate about 1 W each from sensor leakage currents at the end of their lifetime, and the highest-occupancy TRT straws dissipate about 10 mW each at the LHC design luminosity.
- for all of the above reasons, it has been well known since the early 90's in the LHC community that the material budget of the tracker systems as built would pose serious problems in terms of their own performance (see Sect. 16.8.1) and even more so in terms of the intrinsic performance of the electromagnetic calorimeter and of the overall performance for electron/photon measurements (see Sect. 16.8.2). Despite the best efforts of the community, the material budget for the tracker has risen steadily over the years and reached values of two radiation lengths (X_0) and close to 0.6 interaction lengths (λ) in the worst regions (see Sect. 16.3.2.1 for more details and plots).

The high-radiation environment imposes stringent conditions on the inner-detector sensors, on-detector electronics, mechanical structure and services. Over the 10-year design lifetime of the experiment, the pixel inner vertexing layer must be replaced after approximately 3 years of operation at design luminosity. The other pixel layers and the pixel disks must withstand a 1 MeV neutron equivalent fluence F_{neq} [18] of up to $\sim 8 \times 10^{14}\,cm^{-2}$. The innermost parts of the SCT must withstand F_{neq} of up to $2 \times 10^{14}\,cm^{-2}$. To maintain an adequate noise performance after radiation damage, the silicon sensors must be kept at low temperature (approximately -5 to $-10\,°C$) implying coolant temperatures of $\sim -25\,°C$. In contrast, the TRT is designed to operate at room temperature.

The above operating specifications imply requirements on the alignment precision which are summarised in Table 16.7 and which serve as stringent upper limits on the silicon-module build precision, the TRT straw-tube position, and the measured module placement accuracy and stability.
This leads to:

(a) a good construction accuracy with radiation-tolerant materials having adequate detector stability and well understood position reproducibility following repeated cycling between temperatures of -20 and $+20\,°C$, and a temperature uniformity on the structure and module mechanics which minimises thermal distortions;
(b) an ability to monitor the position of the detector elements using charged tracks and, for the SCT, laser interferometric monitoring [19];

Table 16.7 Intrinsic measurement accuracies and mechanical alignment tolerances for the tracker sub-systems, as defined by the performance requirements of the ATLAS experiment

Item	Intrinsic accuracy [μm]	Alignment tolerances [μm]		
		Radial (R)	Axial (z)	Azimuth ($R-\phi$)
Pixel				
Layer-0	10 (R-ϕ) 115 (z)	10	20	7
Layer-1 and Layer-2	10 (R-ϕ) 115 (z)	20	20	7
Disks	10 (R-ϕ) 115 (R)	20	100	7
SCT				
Barrel	17 (R-ϕ) 580 (z)[a]	100	50	12
Disks	17 (R-ϕ) 580 (R)[a]	50	200	12
TRT	130			30[b]

The numbers in the table correspond to the single-module accuracy for the pixels, to the effective single-module accuracy for the SCT and to the drift-time accuracy of a single straw for the TRT
[a]Arises from the 40 mrad stereo angle between back-to-back sensors on the SCT modules with axial (barrel) or radial (end-cap) alignment of one side of the structure. The result is pitch-dependent for end-cap SCT modules
[b]The quoted alignment accuracy is related to the TRT drift-time accuracy

(c) a trade-off between the low material budget needed for optimal performance and the significant material budget resulting from a stable mechanical structure with the services of a highly granular detector.

The design and construction of systems, capable of meeting the physics requirements and of providing stable and robust operation over many years, has been perhaps the most formidable challenge faced by the experiment because of the very harsh radiation conditions to be faced near the interaction point and of the conflicting requirements in terms of material budget between the physics and the design constraints. The latter arise mostly from the on-detector high-speed front-end electronics, which require a lot of power to be fed into a limited volume and therefore a large amount of heat to be removed from a very distributed set of local heat sources across the whole tracker.

This section describes briefly the ATLAS tracker and its main properties and discusses a few salient aspects from the construction experience and from the measured performance in laboratory and test beam of production modules in the various technologies. A few examples of the overall performance expected in the actual configuration of the experiment are presented in Sect. 16.8.1, where it is also compared to the expected performance of the CMS tracker.

16.3.2 Construction Experience

16.3.2.1 General Aspects

The ATLAS tracker system has evolved considerably since the submission of the Technical Proposal in 1994 and even since the corresponding Technical Design Reports in 1997/1998. The evolution was dictated by many factors, some of which have already been alluded to in Sect. 16.2.3 and some of which are related to the specific design challenges posed:

- the rapid development of radiation-hard silicon sensors and of their front-end electronics led many physicists and engineers in the community to focus for a long time on the single module scale and, as a consequence, to perhaps address some of the systems issues, especially for the readout and cooling aspects, too late.
- the legitimate concerns throughout the collaborations about the material budget of the tracker systems resulted in huge pressures on the engineering design effort in terms of materials at a very early stage. This effort has been largely successful in terms of mechanics, as can be seen from the very light and state-of-the-art structures used to support and hold the detector components in the tracker system. The already considerable experience from the space industry across the world turned out to be invaluable, including in terms of thermal behaviour and of resistance to radiation and to moisture absorption.
- the tracker macro-assemblies, once completed as operational devices, are the sum of a large number of diverse and tiny components. Many of these components were not built into the design from the very beginning and only general assumptions based on past experience were made concerning their manufacture. Several of these assumptions turned out to be incorrect: for example, the use of silver in the electrical connections and cables has had to be minimised because of activation issues. The pressure on the material budget led to the choice of risky technical solutions for cooling and power, involving hard-to-validate thin-walled Aluminium, copper/Nickel or Titanium pipes and polyimide/Aluminium tapes rather than the less risky but heavier stainless steel pipes and polyimide/copper tapes.
- many of the systems aspects were discovered as the detailed design progressed, rather than foreseen early on, and this has led to difficult retrofitting exercises and sometimes to technical solutions more complex and risky than those which would be devised from a clean slate today. Some of the substrates for the electronics of the silicon modules barely existed in terms of conceptual design at a time when the front-end electronics chip was ready for production. This is one example of a specific and critical component, which was not always incorporated into the detailed design of the system from the very beginning.

  Another more general example stems from the engineering choices made for the implementation of the on-detector and off-detector cooling systems: there are as many on-detector cooling schemes and pipe material choices as there are

detector components. The cooling systems themselves are all operating under severe space limitations on-detector and at high pressure (from 3 to 6 bars). These systems range from room-temperature monophase C_6F_{14} for the TRT to cold evaporative C_3F_8 for the SCT and pixels. Many problems have been encountered during the commissioning in situ and early operation of these systems, and it is fair to say a posteriori that this is one area where a stronger and more centralised engineering effort would have probably come up with a more uniform, more robust and redundant, and less risky implementation.

- Table 16.8 shows how optimistic the estimate of the material budget of the ATLAS tracker was at the time of the Technical Proposal in 1994 and how it has evolved since then to reach the values quoted in early 2008, after completion of the installation of all of its components. These values cannot be claimed to be final yet, although most of the remaining uncertainties are small and related to the exact routing details of the various services and of patch-panels for cable and pipe connections. These are situated within the tracker volume, but not always in the fiducial region where the detectors expect to perform precision tracking and electromagnetic calorimetry measurements (for example, the patch-panels for the pixel detector are outside this fiducial region). The material budget for the tracker has risen steadily over the years and the only significant decrease seen (from 1997 to now) is due to the rerouting of the pixel services from a large radius along the LAr barrel cryostat to a much smaller radius along the pixel

Table 16.8 Evolution of the amount of material expected in the ATLAS tracker from 1994 to 2007

Date	ATLAS tracker material budget estimate $[X/X_0]$	
	$\lvert\eta\rvert \approx 0$	$\lvert\eta\rvert \approx 1.7$
1994 (Technical Proposal)	0.20	0.70
1997 (Technical Design Report)	0.25	1.50
End 2005 (End of construction)	0.40	1.35
Summer 2007 (End of installation)	0.47	2.40

The numbers are given in fractions of radiation lengths (X/X_0). Note that, for ATLAS, the reduction in material from 1997 to 2006 at $\eta \approx 1.7$ is due to the rerouting of pixel services from an integrated barrel tracker layout with pixel services along the barrel LAr cryostat to an independent pixel layout with pixel services routed at much lower radius and entering a patch panel outside the acceptance of the tracker (this material appears now at $\eta \approx 3$). Note also that the numbers do not represent all the material seen by particles before entering the active part of the electromagnetic calorimeter, since particles see in addition the barrel LAr cryostat and the solenoid coil (amounting to approximately $2X_0$ at $\eta = 0$) or the end-cap LAr cryostat at the larger rapidities

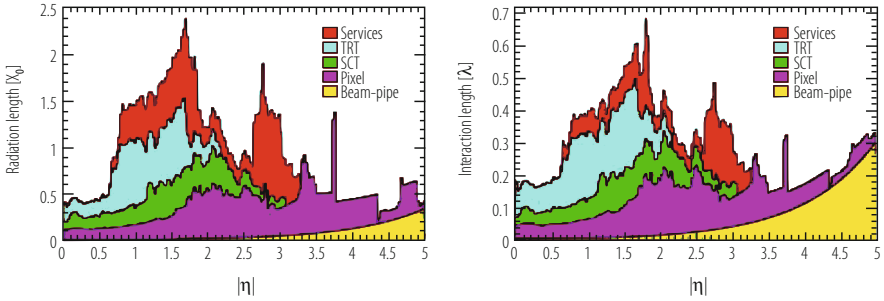


Fig. 16.10 Material distribution (X_0, λ) at the exit of the ATLAS tracker, including the services and thermal enclosures. The distribution is shown as a function of $|\eta|$ and averaged over ϕ. The breakdown indicates the contributions of external services and of individual sub-detectors, including services in their active volume. These plots do not include additional material just in front of the electromagnetic calorimeter, which is quite large in ATLAS (LAr cryostats and, for the barrel, solenoid coil)

support tube, a significant change in the ATLAS tracker design, which occurred in 1999.

Figure 16.10 shows how this material budget is distributed as a function of pseudorapidity. The material closest to the beam (pixel detectors) is clearly the one most critical for the performance of the tracker and of the electromagnetic calorimetry: this amounts to between 10 and 50% X/X_0. The material budget can also be broken down in terms of its functional components: a large contribution to the material budget arises from cooling and cables in areas where these services accumulate to be routed radially outwards, towards the cracks in the electromagnetic calorimetry foreseen for their passage. It is therefore not surprising that, until all the details of the granularity, technical components, routing, fixation schemes, etc., were known and incorporated into assembly drawings and detailed spreadsheets, the material budgets announced for this tracker of unprecedented scope and complexity were largely underestimated.

16.3.2.2 Silicon-Strip and Straw Tube Trackers

The ATLAS SCT contains a total of 4088 modules corresponding to 6.3 million channels, of which 99.7% have been measured to be fully operational in terms of electrical and thermal performance in situ. The ATLAS TRT comprises approximately 350,000 channels, of which about 98.5% fully meet the operational specifications in terms of noise counting rate and of basic efficiency and high-voltage behaviour.

The ATLAS tracker was installed in three successive stages, from summer 2006 (barrel SCT/TRT tracker), to end 2006 (end-cap SCT/TRT trackers), and to spring 2007 (pixels). It is impossible to properly give credit here to all the work performed

over the past 15 years to validate the design choices involving each and every one of the delicate components composing these tracking detectors. Only a few of the most prominent examples are quoted below:

- all the front-end electronic designs had to be submitted to stringent specifications in terms of survival to very high ionisation doses and neutron fluences and of robustness against single-event upsets. The performance of fully irradiated and operational modules equipped with the latest iteration in the design had to be repeatedly measured and characterised in laboratory tests and particle beams of various types and intensities [20].
- each component in contact with the active gas of the ATLAS TRT straws has had to be validated in a well-controlled set-up over many hundreds of hours of accelerated ageing tests using the gas mixture chosen for operation in the experiment. This was necessary because impurities of only a few parts per billion, picked up somewhere in the system, could be deposited on the wires and thereby destroy the gas gain in an irrecoverable way [21]. One critical component in the barrel TRT modules, a glass bead serving as wire joint to separate the two halves of each wire, actually failed the ageing tests with the originally chosen gas mixture ($Xe–CO_2–CF_4$) and the collaboration had to eventually change the gas mixture to the current one ($Xe–CO_2–O_2$), in which the fluorine component has been removed. This gas mixture reduces the direct risk to the wire joints, but is somewhat less stable operationally and does not have the same self-cleaning properties as the original one.

16.3.2.3 Pixel Detectors

The ATLAS pixel detector has been one of the last elements installed in the experiment, in great part for practical reasons, but also because this is the detector which has undergone the most difficult development path. It can perhaps be considered as the most striking example of the marvels achieved during the long and painstaking years of research and development: the pixel detector will survive over many years in the most hostile region of the experiment and deliver some of the most important data required to understand in detail what will be happening within a few tens of microns from the interaction point.

Fifteen years ago, at the time of the ATLAS Technical Proposal, very few physicists believed that these detectors could be built within the specifications required in terms of radiation hardness and of readout bandwidth and speed. Today, the data collected using cosmic rays (in 2008 and 2009) and early collisions (end of 2009) have demonstrated that the pixel detector works as expected. The future will tell how long the innermost layer will survive, but the collaboration is already proceeding towards a strategy of "replacement" of the innermost pixel layer on the timescale of 2015. This innermost layer is not expected to survive over the full time-span of the operation of the experiment, which should lead to integrated luminosities

Table 16.9 Main parameters of the ATLAS pixel system

Number of hits per track	3
Total number of channels	$80 \cdot 10^6$
Pixel size (μm in $R\phi$) $\times$ (μm in z/R)	50×400
Lorentz angle [degrees], initial...end	12...4
Tilt in $R\phi$ [degrees]	20 (only barrel)
Total active area of silicon [m^2]	1.7 (n^+/n)
Sensor thickness [μm]	250
Total number of modules	1744 (288 in disks)
Barrel layer radii [mm]	50.5, 88.5, 122.5
Disk layer min.... max. radii [mm]	88.8...149.6
Disk positions in z [cm]	49.5, 58.0, 65.0
Signal-to-noise ratio for m.i.p. (initial)	120
Total fluence at $L = 10^{34}$ (n_{eq}/cm^2/year) at radius of 5 cm (innermost layer)	$3 \cdot 10^{14}$
Signal-to-noise ratio (after 10^{15} n_{eq}/cm^2)	80
Resolution in $R\phi$ (μm)	$\approx$10
Resolution in z/R (μm)	$\approx$100

of close to 300 fb^{-1}. Table 16.9 shows the most relevant parameters concerning the ATLAS pixel system.

Finally, Fig. 16.11 shows the results of test-beam measurements of the $R\phi$ accuracy of production modules of the ATLAS pixel detector before and after being irradiated with a total equivalent fluence corresponding to about 10^{15} neutrons per cm^2 [22]. These results are somewhat optimistic because they were obtained with analogue readout and at an ideal incidence angle, but they nevertheless demonstrate the extreme robustness of the pixel modules constructed for ATLAS. This is one striking example of the painstaking validation work done in the early phase of the construction years.

16.4 Calorimeter System

The design of the ATLAS calorimeter system is to a large extent the end product of about 25 years of development and experience gained over several generations of high-energy colliders and general-purpose experiments, all of which have brought major advances in the understanding of the field. These advances range from the concept of full coverage in total transverse energy at UA1, to that of precision hadron calorimetry at ZEUS, and to that of very high granularity of the electromagnetic calorimeters and the use of energy-flow techniques in the LEP detectors [23].

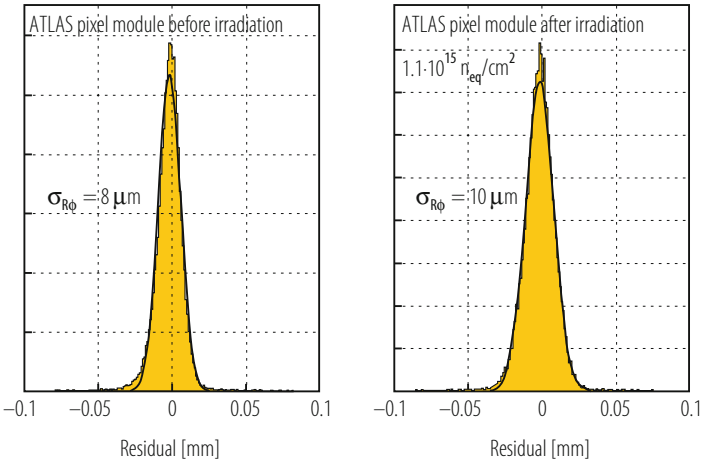


Fig. 16.11 Residuals from $R\phi$ measurements of production-grade ATLAS pixel module before irradiation (left) and after being irradiated with a total equivalent fluence corresponding to about 10^{15} neutrons per cm^2 (right), as obtained from test-beam data taken in 2004. The contribution of the track extrapolation to the width of the residuals is about 5 μm (it should be subtracted in quadrature from the overall residual widths quoted in the figure to obtain the intrinsic resolution of the tested module)

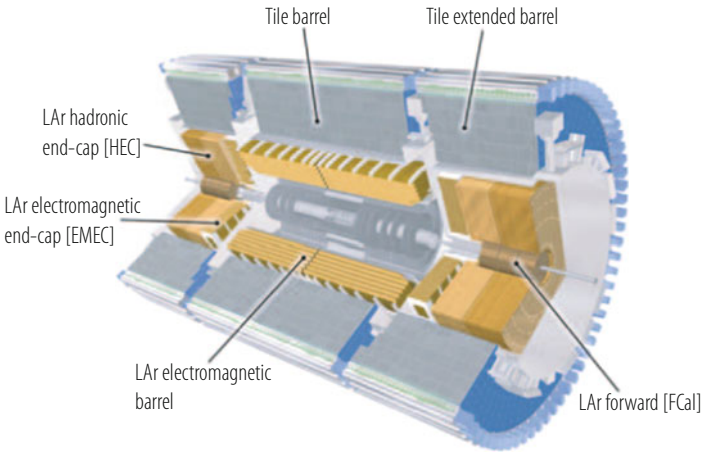


Fig. 16.12 Cut-away view of the ATLAS calorimeter system. The various calorimeter components are clearly visible, from the LAr barrel and end-cap electromagnetic calorimeters, to the scintillating tile barrel and extended barrel hadronic calorimeters, and to the LAr end-cap and forward hadronic calorimeters

The ATLAS calorimeter system, as depicted overall in Fig. 16.12, will play a crucial role at the LHC for two main reasons: first, its intrinsic resolution improves with energy, in contrast to magnetic spectrometers; second, it will provide the trigger primitives for all the high-p_T objects of interest to the experiments except for the muons.

The integration of a hermetic and high-precision calorimeter system into the overall design of the ATLAS detector and its magnet systems has been a task of high complexity where compromises have had to be made, as will be shown in the first part of this section, which describes the basic requirements and features of the calorimeters. As illustrated in the second part, which highlights some aspects of the construction of the most critical element, namely the electromagnetic calorimeter, and of its measured performance in test beam, the impact of the main design choices and of the technology implementations on the performance has been very significant. A few examples of the overall performance expected in the actual configuration of the experiment are presented in Sect. 16.8.2, where it is also compared to the expected performance of the CMS calorimeter system.

16.4.1 General Considerations

16.4.1.1 Performance Requirements

The main performance requirements from the physics on the calorimeter system can be briefly summarised as follows:

- excellent energy and position resolution together with powerful particle identification for electrons and photons within the relevant geometrical acceptance (full azimuthal coverage over $|\eta| < 2.5$) and over the relevant energy range (from a few GeV to several TeV). The electron and photon identification requirements are particularly demanding at the LHC, as already explained in Sect. 16.2.1. These considerations induce requirements of high granularity and low noise on the calorimeters. One has to add to this the operational requirements of speed of response and resistance to radiation (the electromagnetic calorimeters will have to withstand neutron fluences of up to 10^{15} n/cm^2 and ionising radiation doses of up to 200 kGy over 10 years of LHC operation at design luminosity).
- excellent jet energy resolution within the relevant geometrical acceptance, which is similar to that foreseen for the electron and photon measurements (see above). The quality of the jet energy resolution would play an important role in the case of discovery of supersymmetric particles with cascade decays into many hadronic jets [24].
- good jet energy measurements over the coverage required to contain the full transverse energy produced in hard-scattering collisions at the LHC. A calorimetry coverage over $|\eta| < 5$ is necessary to unambiguously ascribe the observation of significant missing transverse energy to non-interacting particles, such as neutrinos from W-boson decay or light neutralinos from supersymmetric particle cascade decays. With adequate calorimetry coverage providing precise measurements of the missing transverse energy, the experiments will be able to reconstruct invariant masses of pairs of hadronically decaying τ-leptons produced for example in the decays of supersymmetric Higgs bosons. They

will also thus be able to identify forward jets produced in vector-boson fusion processes.
- good separation between hadronic showers from QCD jets and those from decays of τ-leptons.
- fast and efficient identification of the processes of interest at the various trigger levels, in particular for the L1 trigger (see Sect. 16.6).

16.4.1.2 General Features of Electromagnetic Calorimetry

The ATLAS EM calorimeter [25] is divided into a barrel part covering approximately $|\eta| < 1.5$ and two end-caps covering $1.4 < |\eta| < 3.2$, and its main parameters are listed in Table 16.10. Its fiducial coverage is without appreciable cracks, except in the transition region between the barrel and end-cap cryostats, where the measurement accuracy is degraded over $1.37 < |\eta| < 1.52$ because of large energy losses in the material in front of the active EM calorimeter, which reaches up to $6\,X_0$. The excellent uniformity of coverage is a direct consequence of the design of this lead/liquid Argon sampling calorimeter with accordion-shaped electrodes and absorbers. The total thickness of the EM calorimeter varies from a minimum of $24\,X_0$ (at $\eta \approx 0$) to a maximum of $35\,X_0$ (at $\eta \approx 2.5$). This depth is sufficient to contain EM showers at the highest energies (a few TeV) and preserve the energy resolution, in particular the constant term which is dominant above a few hundred GeV.

As can be seen from Table 16.10, the ATLAS EM calorimeter has been designed with both excellent lateral and longitudinal granularity, with samplings in depth optimised for energy loss corrections (presampler) and for shower pointing accuracy together with γ/π^0 and electron/jet separation (strips). The intrinsic performance of the EM calorimeter is however significantly affected by the unavoidable amount of material which had to be incorporated in the tracker system (see Fig. 16.10), and also by the cryostats and the solenoid coil in the case of the ATLAS EM calorimeter (see Sect. 16.8.2 for more details).

16.4.1.3 General Features of Hadronic Calorimetry

Figure 16.13 shows the total number of absorption lengths contained in the ATLAS hadronic calorimetry and in front of the muon system as a function of pseudorapidity. Good containment of jets of typically 1 TeV energy requires about $11\,\lambda$ in the full calorimeter, a target which has been achieved over most of the pseudorapidity range.

For the central part of the hadronic calorimetry, which covers the range $0 < |\eta| < 1.7$, the sampling medium consists of scintillator tiles and the absorber medium is steel. The tile calorimeter is composed of three parts, one central barrel and two extended barrels. The choice of this technology provides maximum radial depth for the least cost for ATLAS. The hadronic calorimetry is extended to

Table 16.10 Main parameters of the ATLAS calorimeter system

	Barrel		End-cap	
EM calorimeter				
Number of layers and $\|\eta\|$ coverage				
Presampler	1	$\|\eta\| < 1.52$	1	$1.5 < \|\eta\| < 1.8$
Calorimeter	3	$\|\eta\| < 1.35$	2	$1.375 < \|\eta\| < 1.5$
	2	$1.35 < \|\eta\| < 1.475$	3	$1.5 < \|\eta\| < 2.5$
			2	$2.5 < \|\eta\| < 3.2$
Granularity $\Delta\eta \times \Delta\phi$ versus $\|\eta\|$				
Presampler	0.025×0.1	$\|\eta\| < 1.52$	0.025×0.1	$1.5 < \|\eta\| < 1.8$
Calorimeter (strip layer)	$0.025/8 \times 0.1$	$\|\eta\| < 1.40$	0.050×0.1	$1.375 < \|\eta\| < 1.425$
	0.025×0.025	$1.40 < \|\eta\| < 1.475$	0.025×0.1	$1.425 < \|\eta\| < 1.5$
			$0.025/8 \times 0.1$	$1.5 < \|\eta\| < 1.8$
			$0.025/6 \times 0.1$	$1.8 < \|\eta\| < 2.0$
			$0.025/4 \times 0.1$	$2.0 < \|\eta\| < 2.4$
			0.025×0.1	$2.4 < \|\eta\| < 2.5$
			0.1×0.1	$2.5 < \|\eta\| < 3.2$
Calorimeter (middle layer)	0.025×0.025	$\|\eta\| < 1.40$	0.050×0.025	$1.375 < \|\eta\| < 1.425$
	0.075×0.025	$1.40 < \|\eta\| < 1.475$	0.025×0.025	$1.425 < \|\eta\| < 2.5$
			0.1×0.1	$2.5 < \|\eta\| < 3.2$
Calorimeter (back layer)	0.050×0.025	$\|\eta\| < 1.35$	0.050×0.025	$1.5 < \|\eta\| < 2.5$
Number of readout channels				
Presampler	7808		1536 (both sides)	
Calorimeter	101,760		62,208 (both sides)	
LAr hadronic end-cap				
$\|\eta\|$ coverage			$1.5 < \|\eta\| < 3.2$	
Number of layers			4	
Granularity $\Delta\eta \times \Delta\phi$			0.1×0.1	$1.5 < \|\eta\| < 2.5$
			0.2×0.2	$2.5 < \|\eta\| < 3.2$
Readout channels			5632 (both sides)	
LAr forward calorimeter				
$\|\eta\|$ coverage			$3.1 < \|\eta\| < 4.9$	
Number of layers			3	

(continued)

Table 16.10 (continued)

	Barrel		End-cap	
Granularity $\Delta x \times \Delta y$ [cm]			FCal1: 3.0 × 2.6	$3.15 < \|\eta\| < 4.30$
			FCal1: ∼ four times finer	$3.10 < \|\eta\| < 3.15$,
				$4.30 < \|\eta\| < 4.83$
			FCal2: 3.3 × 4.2	$3.24 < \|\eta\| < 4.50$
			FCal2: ∼ four times finer	$3.20 < \|\eta\| < 3.24$,
				$4.50 < \|\eta\| < 4.81$
			FCal3: 5.4 × 4.7	$3.32 < \|\eta\| < 4.60$
			FCal3: ∼ four times finer	$3.29 < \|\eta\| < 3.32$,
				$4.60 < \|\eta\| < 4.75$
Readout channels			3524 (both sides)	
Scintillator tile calorimeter				
	Barrel		Extended barrel	
$\|\eta\|$ coverage	$\|\eta\| < 1.0$		$0.8 < \|\eta\| < 1.7$	
Number of layers	3		3	
Granularity $\Delta\eta \times \Delta\phi$	0.1 × 0.1		0.1 × 0.1	
Last layer	0.2 × 0.1		0.2 × 0.1	
Readout channels	5760		4092 (both sides)	

larger pseudorapidities by a copper/liquid-argon calorimeter system, which covers the range $1.5 < |\eta| < 3.2$, and by the forward calorimeters, a set of copper-tungsten/liquid-argon detectors at larger pseudorapidities. The hadronic calorimetry thus reaches one of its main design goals, namely coverage over $|\eta| < 4.9$.

The ATLAS forward calorimeters are fully integrated into the cryostat housing the end-cap calorimeters, which reduces the neutron fluence in the muon system and, with careful design, affects very little the neutron fluence in the tracker volume. The main role of these calorimeters is to keep the tails in the measurement of missing transverse energy at a low level and to tag jets in the forward direction rather than to accurately measure their energy, so their geometry has been simplified and their readout costs have been minimised. The forward calorimeters are based on copper (front) and tungsten (back) absorber bodies and absorber rods, the latter being parallel to the beam and slotted into precisely machined holes. The gaps in these holes are filled with LAr and operated at an electric field of about 1 kV/mm.

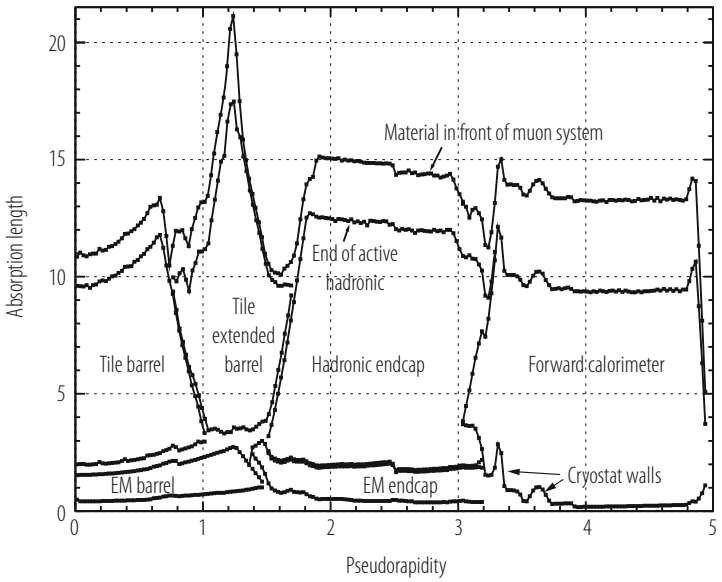


Fig. 16.13 Distribution of amount of material (in absorption lengths) for the ATLAS calorimetry (and in front of the muon system) as a function of η

16.4.2 Construction Experience and Measured Performance in Test Beam

As has been described above, the ATLAS calorimeters comprise a variety of technologies, each with its own challenges and pitfalls, and only a few of the most prominent examples of lessons learned during construction can be given in this review.

The biggest challenge has clearly been the construction of the electromagnetic calorimeters. The technology chosen for the ATLAS EM calorimeter, although based on a well established technique had a number of innovative features, which resulted in some major production issues:

- the most difficult part of the project, by far, has been the fabrication in industry of large electrodes of about 2 m length containing about 1000 resistive pads each. This problem was overcome through the careful monitoring of the production on-site by experts from the collaboration.
- a total of about 20,000 m^2 of honeycomb spacers have been used to maintain the flexible electrodes in the centre of the gap between absorbers. To avoid major problems with the high-voltage behaviour of assembled modules, a rigorous and careful cleaning procedure for all parts, especially the honeycomb, had to be implemented.

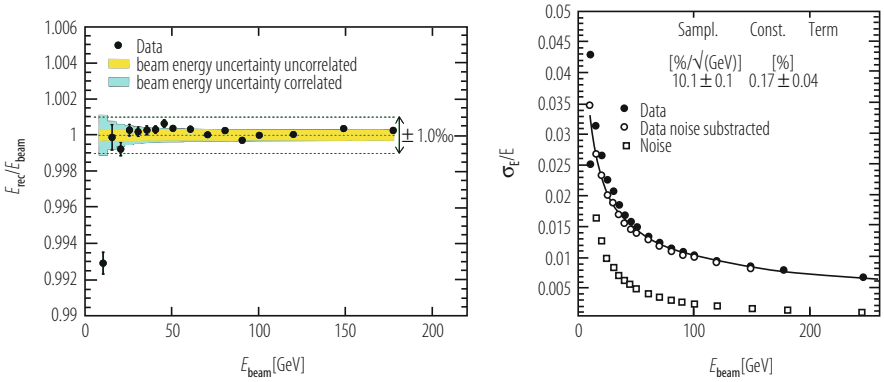


Fig. 16.14 Linearity of response (left) and energy resolution (right) obtained for a production module of the ATLAS barrel EM calorimeter as a function of the incident electron beam energy

- radiation-tolerant electronics had to be produced for all components in the cavern. This comprises all the front-end electronics boards housed near the signal feed-throughs.

The ATLAS collaboration has performed an extensive programme of test-beam measurements to calibrate and characterise the EM calorimeter modules [26]. The original plans called for a test-beam calibration of about 20% of the modules. In the end, a smaller fraction of 15% of the ATLAS EM modules underwent detailed test-beam measurements, and a few recent results from these stand-alone calibration campaigns are presented here.

Figure 16.14 shows that a linearity of response of ±1 per mil has been obtained over an electron energy range from 20 to 180 GeV for an ATLAS barrel LAr EM module. To achieve this, while preserving the energy resolution (also shown in Fig. 16.14), requires a thorough understanding of the material in front of the active calorimeter and a careful evaluation of the weights and corrections to be applied to the raw cluster energy. The uniformity of response across the whole module has also been measured and found to contribute an r.m.s. of 0.4% to the global constant term, which is within the specifications set to the LAr EM calorimeter (see Sect. 16.8.2 for a more detailed discussion of the various contributions to the constant term for the EM calorimeters).

16.5 Muon Spectrometer System

Muons are a very robust, clean and unambiguous signature of much of the physics that ATLAS has been designed to study. The ability to trigger and to reconstruct muons at the highest luminosities of the LHC has been incorporated into the design of the experiment from the very beginning [29]. In fact, the concepts chosen for

measuring muon momenta have shaped the experiment more than any other physics consideration (see also Sect. 16.2.1).

As discussed already in Sect. 16.2.2, the choice of magnet was motivated by the method which would be used for the measurement of muons with momenta up to $\sim$TeV scales. ATLAS has thus opted for a high-resolution, stand-alone measurement independently of the rest of the sub-detectors, resulting in a very large volume, with low material density, over which the muon measurement takes place. The ATLAS toroidal magnetic field provides a momentum resolution which is essentially independent of pseudorapidity up to a value of 2.7.

This section reviews the main features of the muon spectrometer system and discusses a few of the challenges encountered. A few examples of the overall performance expected in the actual configuration of the experiment are presented in Sect. 16.8.3, where it is also compared to the expected performance of the CMS muon system.

16.5.1 General Considerations

The physics signatures that give rise to muons are numerous and varied. At the highest momenta, they include muons from new high-mass (multi-TeV) resonances such as heavy neutral gauge bosons, Z', as well as decays from heavy Higgs bosons. At the lowest end of the spectrum, B-physics relies on the reconstruction of muons with momentum down to a few GeV. The resulting requirements are:

- Resolution: the 'golden' decay of the Standard Model Higgs boson into four muons, $H \rightarrow ZZ \rightarrow 4\,\mu$, requires the ability to reconstruct the momentum and thus mass of a narrow two-muon state with a precision at the level of 1%. At the upper end of the spectrum, the goal is to achieve a 10% momentum resolution for 1 TeV muons.
- Wide rapidity coverage: almost two-thirds of the decays of an intermediate-mass Higgs boson to four muons have at least one muon in the region $|\eta| > 1.4$. A hermetic system, which measures muons up to $|\eta| \sim 2.5$, has turned out to be the best compromise.
- Identification inside dense environments, e.g. hadronic jets or regions with high backgrounds.
- Trigger: the ability to measure the momenta of muons online on a stand-alone basis, i.e. without reference to any other detector system, and to select events with muons above 5–10 GeV momentum is of paramount importance.

There are also the requirements which result from the 25 ns spacing in time between successive beam crossings and from the neutron radiation environment of the experimental halls. Good timing resolution and the ability to identify the bunch-crossing in question, as well as redundancy in the measurements, are therefore also demanded of the muon detectors, which represent by far the largest and most difficult system to install in the experiment.

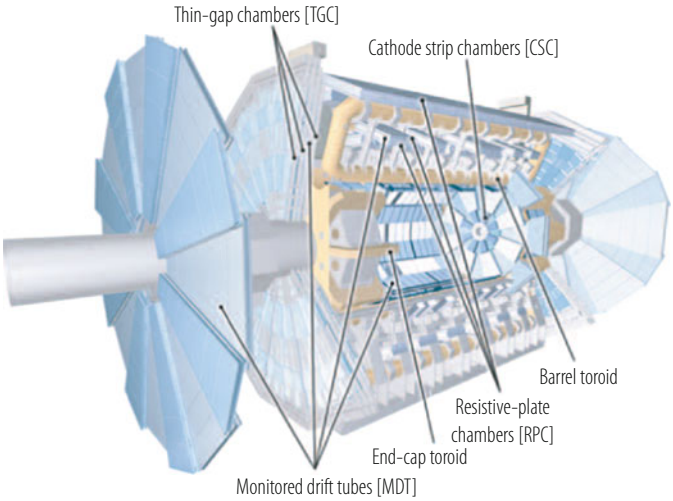


Fig. 16.15 Cut-away view of the ATLAS muon spectrometer system. displaying the regions in which the different muon chamber technologies are used

The conceptual layout of the muon spectrometer is shown in Fig. 16.15 and the main parameters of the muon chambers are listed in Table 16.11. It is based on the magnetic deflection of muon tracks in the large superconducting air-core toroid magnets, instrumented with separate trigger and high-precision tracking chambers. Over the range $|\eta| < 1.4$, magnetic bending is provided by the large barrel toroid. For $1.6 < |\eta| < 2.7$, muon tracks are bent by two smaller end-cap magnets inserted into both ends of the barrel toroid. Over $1.4 < |\eta| < 1.6$, usually referred to as the transition region, magnetic deflection is provided by a combination of barrel and end-cap fields. This magnet configuration provides a field which is mostly orthogonal to the muon trajectories, while minimising the degradation of resolution due to multiple scattering. The anticipated high level of particle flux has had a major impact on the choice and design of the spectrometer instrumentation, affecting performance parameters such as rate capability, granularity, ageing properties, and radiation hardness. In the barrel region, tracks are measured in chambers arranged in three cylindrical layers around the beam axis; in the transition and end-cap regions, the chambers are installed in planes perpendicular to the beam, also in three layers.

16.5.1.1 Muon Chamber Types

Over most of the η-range, a precision measurement of the track coordinates in the principal bending direction of the magnetic field is provided by Monitored Drift Tubes (MDT's). The mechanical isolation in the drift tubes of each sense wire from its neighbours guarantees a robust and reliable operation. At large pseudorapidities, Cathode Strip Chambers (CSC's, which are multiwire proportional chambers with

Table 16.11 Main parameters of the ATLAS muon spectrometer

Monitored drift tubes	MDT
Coverage	$\|\eta\| < 2.7$ (innermost layer: $\|\eta\| < 2.0$)
Number of chambers	1088 (1150)
Number of channels	339,000 (354,000)
Function	Precision tracking
Cathode strip chambers	CSC
Coverage	$2.0 < \|\eta\| < 2.7$
Number of chambers	32
Number of channels	31,000
Function	Precision tracking
Resistive plate chambers	RPC
Coverage	$\|\eta\| < 1.05$
Number of chambers	544 (606)
Number of channels	359,000 (373,000)
Function	Triggering, second coordinate
Thin gap chambers	TGC
Coverage	$1.05 < \|\eta\| < 2.7$ (2.4 for triggering)
Number of chambers	3588
Number of channels	318,000
Function	Triggering, second coordinate

Numbers in brackets for the MDT's and the RPC's refer to the final configuration of the detector in 2009

cathodes segmented into strips) with higher granularity are used in the innermost plane over $2 < |\eta| < 2.7$, to withstand the demanding rate and background conditions. The stringent requirements on the relative alignment of the muon chamber layers are met by the combination of precision mechanical-assembly techniques and optical alignment systems both within and between muon chambers.

The trigger system covers the pseudorapidity range $|\eta| < 2.4$. Resistive Plate Chambers (RPC's) are used in the barrel and Thin Gap Chambers (TGC's) in the end-cap regions. The trigger chambers for the muon spectrometer serve a threefold purpose: provide bunch-crossing identification, provide well-defined p_T thresholds, and measure the muon coordinate in the direction orthogonal to that determined by the precision-tracking chambers.

16.5.1.2 Muon Chamber Alignment and B-Field Reconstruction

The overall performance over the large areas involved, particularly at the highest momenta, depends on the alignment of the muon chambers with respect to each other and with respect to the overall detector.

The accuracy of the stand-alone muon momentum measurement necessitates a precision of 30 μm on the relative alignment of chambers both within each projective tower and between consecutive layers in immediately adjacent towers. The internal deformations and relative positions of the MDT chambers are monitored by approximately 12,000 precision-mounted alignment sensors, all based on the optical monitoring of deviations from straight lines. Because of geometrical constraints, the reconstruction and/or monitoring of the chamber positions rely on somewhat different strategies and sensor types in the end-cap and barrel regions, respectively.

The accuracy required for the relative positioning of non-adjacent towers to obtain adequate mass resolution for multi-muon final states, lies in the few millimetre range. This initial positioning accuracy is approximately established during the installation of the chambers. Ultimately, the relative alignment of the barrel and forward regions of the muon spectrometer, of the calorimeters and of the tracker will rely on high-momentum muon trajectories.

For magnetic field reconstruction, the goal is to determine the bending power along the muon trajectory to a few parts in a thousand. The field is continuously monitored by a total of approximately 1800 Hall sensors distributed throughout the spectrometer volume. Their readings are compared with magnetic-field simulations and used for reconstructing the position of the toroid coils in space, as well as to account for magnetic perturbations induced by the tile calorimeter and other nearby metallic structures.

The muon system consists of three large superconducting air-core toroid magnets, which are instrumented with different types of chambers to provide the two needed functions, namely high-precision tracking and triggering. The central (or barrel) region, $|\eta| < 1.0$, is covered by a large barrel magnet consisting of eight coils which surround the hadron calorimeter. In this region, tracks are measured in chambers arranged in three cylindrical layers (stations) around the beam axis. In the end-cap region, $1.4 < |\eta| < 2.7$, muon tracks are bent in two smaller end-cap magnets inserted into both ends of the barrel toroid. The intermediate (transition) region, $1.0 < |\eta| < 1.4$, is less straightforward, since here the barrel and end-cap fields overlap, thus partially reducing the bending power. To keep a uniform resolution in this region, tracking chambers are place in strategic places to improve the quality and accuracy of the measurement. Due to financial constraints, one out of three sets of chambers in this region has been staged, thus leading to an inferior performance in the transition region for the first years of data-taking.

The layout of the ATLAS muon spectrometer system is shown in Fig. 16.15. A total of four types of detectors are used, the choice of technology being driven by the very large surface to be covered, by trigger and precision measurement requirements, and by the different radiation environments. Resistive Plate Chambers (RPC) in the barrel region ($|\eta| < 1.05$) and Thin Gap Chambers (TGC) in the end-cap regions ($1.05 < |\eta| < 2.4$) are used for triggering purposes. These chambers provide a fast response with good time resolution but rather coarse position resolution. The precision measurements are performed by Monitored Drift Tubes (MDT) over most of the coverage. In the regions at large $|\eta|$, where background

conditions are harsher and the rate of muon hits is therefore larger, Cathode Strip Chambers (CSC) are used.

The basic principle of the muon measurement in the ATLAS muon spectrometer is to obtain three segments (or super-points) along the muon trajectory. For momenta up to 300 GeV, the resolution is limited to a few percent by multiple scattering and fluctuations in the energy loss in the calorimeters, and can therefore be improved by combining the momentum measurement with that obtained in the Inner Detector. The momentum resolution goals quoted above at higher momenta imply a very high precision of 80 μm on the individual hits, given the three-point measurement and the available bending power. The required precision on the muon momentum measurement also implies excellent knowledge of the magnetic field. The air-core toroid design leads to a magnetic field, which is modest in average magnitude (0.5 T), but is also inhomogeneous, and must therefore be measured and monitored with high precision (at the level of 20 G). The inhomogeneity of the field and its rapid variations cannot be approximated by simple analytical descriptions and have to be accounted for carefully, thereby enhancing the importance of the use of the inner detector information to reconstruct low-momentum muon tracks with low fake rates.

16.5.1.3 Alignment

Alignment of the muon chambers with respect to each other and with respect to the overall detector is a critical ingredient, key to obtaining the desired performance over the large areas involved, particularly at the highest momenta. The high accuracy of the ATLAS stand-alone measurement necessitates a very high precision of 30 μm on the alignment.

The chambers have however been installed with an accuracy of a few mm, and obviously, no attempt at repositioning the chambers once their installation is completed can realistically be made. Instead, intricate hardware systems have been designed to measure the relative positions between chambers contributing to the measurement of the same tracks, but also to monitor any displacements during the detector operation. These systems are designed to provide continuous monitoring of the positions of the chambers with or without collisions in the accelerator. The very strict requirement of a 30 μm alignment has necessitated the design of a complex system, in which optical sensors are mounted with very high mechanical mounting precision (better than 20 μm in the precise coordinate). The system uses $\sim$5000 alignment sensors, which are either installed on the chambers or in the so-called alignment bars (long instrumented Aluminium cylinders with deformations monitored to within 10 μm, which constitute the alignment reference system in the end-caps). In addition, 1789 magnetic field sensors (3D Hall probes) are also being installed on the chambers to determine with high accuracy the position and shape of the conductors of each coil. From these accurate measurements, the field will be determined throughout the whole volume to an accuracy of about 20 G, provided all magnetic materials are also mapped and described accurately.

The final alignment values will clearly be obtained with the large statistics of muon tracks traversing the muon chambers (rates of about 10 kHz are expected at a luminosity of $10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ for muons with $p_T > 6\,\mathrm{GeV}$).

16.5.2 Construction Experience and Measured Performance in Laboratory and Test Beam

The muon chambers are based on technologies, which were used in previous experiments: drift tubes and CSCs have been used widely in the past; RPCs were used in the L3 and Babar experiments, while TGCs were used in OPAL. Nevertheless, large R&D efforts have been necessary to address the special requirements of the LHC environment.

The high particle fluxes (mainly photons and neutrons) have necessitated searches for the right type of materials and gases, which prevent wire deposits in the case of drift tubes, while new operational modes were developed for the RPCs (proportional regime instead of the streamer regime used in previous experiments) and the TGCs (quasi-proportional mode instead of saturated mode), with the corresponding required changes in the front-end electronics.

In the case of the ATLAS muon spectrometer, the requirement of a precise stand-alone measurement limits the amount of material in order to minimise multiple scattering. This has led to the development of thin but precise Aluminium tubes, which are mounted on very light structures. The deformations of these structures can be monitored by a sophisticated alignment system, as well as the extensive use of paper honeycomb in the trigger chambers to limit the contribution of the detectors in the material description.

Beyond this, the greatest challenge came mostly from the very large, unprecedented areas that the muon chambers had to cover and the correspondingly large numbers of electronic channels. The ATLAS muon system contains approximately $25{,}000\,\mathrm{m}^2$ of active detection planes, and roughly one million electronic channels. The main parameters of the muon chambers are listed in Table 16.11.

The requirement of achieving all this within 'reasonable cost' was actually one of the biggest issues encountered. In terms of lessons learned from the construction process; beyond the general observations made in Sect. 16.2.3, three issues emerge as the most important ones:

- Putting in place, right from the beginning, very tight procedures for quality assurance/quality control (QA/QC). Given the enormous number of elements (wires, strips, tubes, supports) involved, the presence of well-defined and complete QA/QC systems was of the utmost importance. Any and all issues which went unnoticed sooner or later resulted in time and energy-consuming corrective procedures being taken.
- Planning for services. Despite all initial designs and tolerances and safety factors, the cabling procedures always turn out to be more complicated, more time-

consuming and eventually more space-consuming than planned. Whereas the first two issues can, at least in principle, be solved with additional manpower and increased costs, the space issue is a major one, which needs adequate planning right from the start. The space issue has been compounded by the fact that the muon system is traversed by the services of the other detectors, leading to issues of ownership of space and to problems in collecting all the necessary information for proper planning. This major complexity of the actual installation of the services has been one of the major challenges of the Technical Coordination team.

- Uniformity of technologies, power supplies and electronics. As already explained in the introduction, the size of the muon project has necessitated the distribution of the design and construction across different institutes and funding agencies. This necessarily leads to a multitude of different choices for numerous components, from the choice of high-voltage power supplies to basic choices of electronics (ASICs or FPGAs). A strong electronics coordination team is needed to alleviate many of these pressures and lead to an overall system, which will be much easier to maintain.

As for the other detector systems, the ATLAS collaboration has invested a major effort into the validation of the muon spectrometer concept using high-energy test-beam muons. The ATLAS muon test-beam setup had both trigger and tracking chambers placed in the appropriate geometrical positions and equipped with alignment sensors. The most prominent goal (in 2004) was to test the ability to monitor chamber movements and long-term deformations over time-scales of several weeks with the required accuracy, a crucial ingredient for the ultimate accuracy of muon measurements in the TeV range. The test-beam setup included the calculation of deviations from the nominal chamber positions and the storage of the results in a database. These constants were also directly determined by the reconstruction program. The variation of the sagitta as reconstructed in the muon beam, along with that measured from the optical alignment system, was studied over a period covering the thermal fluctuations of a day–night cycle. The spread of the difference between the two distributions was measured to be below 10 μm, i.e. well within the specification of 30 μm. Finally, the correct performance of the trigger was tested with the final trigger electronics prototypes and with all muon systems taking data simultaneously at 40 MHz.

16.6 Trigger and Data Acquisition System

This section briefly describes the main design features and architecture of the ATLAS trigger and data acquisition systems. A few examples of the overall trigger performance expected in the actual configuration of the experiment are presented in Sect. 16.8.4, where it is also compared to the expected performance of the CMS trigger system.

The trigger and data acquisition (DAQ) system of an experiment at a hadron collider plays an essential role because both the collision and the overall data rates are much higher than the rate at which one can write data to mass storage. As mentioned previously, at the LHC, with the beam crossing frequency of 40 MHz, at the design luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, each crossing results in an average of $\sim$23 inelastic p-p collisions with each event producing approximately 1–2 MB of zero-suppressed data. These figures are many orders of magnitude larger than the archival storage as well as the offline processing capability, which correspond to data rates of 200–300 MB/s, or of 100–200 Hz.

The required event rejection power of the real-time system at design luminosity is thus of O(10^7), which is too large to be achieved in a single processing step, if a high efficiency is to be maintained for the physics phenomena of interest. For this reason, the selection task is split into a first, very fast selection step, followed by two steps in which the selection is refined.

The first step (L1 trigger) makes an initial selection based on information of reduced granularity and resolution from only a subset of detectors. This L1 trigger is designed to reduce the rate of events accepted for further processing to less than 100 kHz, i.e. it provides a rejection of a factor $\sim 10^4$ with respect to the collision rate. The figure of 100 kHz is an 'asymptotic' one, to be fully used at the highest luminosities when the beam and experiment conditions demand it, and financial resources allow it. It is expected that at startup, and also during the first years of LHC operation, the L1 trigger will operate at lower rates.

The second step (high-level trigger or HLT) is designed to reduce the L1 accept rate to the final output rate of $\sim 10^2$ Hz. Filtering in the HLT is provided by software algorithms running in large farms of commercial processors, connected to the detector readout system via commercial networks. The physical implementation of the HLT selection is implemented in a two-step process, with independent farms for each of the two steps.

Some key requirements on the overall system are:

- To provide enough bandwidth and computing resources, within financial constraints, to minimise the dead-time at any luminosity, while maintaining the maximum possible efficiency for the discovery signals. The current goal is to have a total dead-time of less than a few (1–2)%. Most of this dead-time is currently planned to occur in the L1 trigger.
- To be robust, i.e. provide an operational efficiency which does not depend significantly on the noise and other conditions in the detector or on changes with time of the calibration and detector alignment constants.
- To provide the possibility of validating and of computing the overall selection efficiencies using only the data themselves, with as little reference to simulation as possible. This implies usage of multiple trigger requirements with overlapping thresholds.
- To uniquely identify the bunch crossing that gave rise to the trigger.
- To allow for the readout, processing and storage of events that will be needed for calibration purposes.

16.6.1 General Considerations

The most important architectural decision in the Trigger/DAQ system is the number of physical entities, or trigger levels, which will be used to provide the rate reduction of $O(10^3)$ from the rate of 100 kHz accepted by the L1 trigger to the final rate to storage of $O(10^2)$ Hz. Current practice for large general-purpose experiments operating at CERN, DESY, Fermilab, KEK and SLAC is to use at least two more entities, colloquially referred to as the L2 and L3 triggers. Some experiments even have a L4 trigger. The higher the level, the more general-purpose the implementation, with the L3 and L4 trigger systems always relying on farms of standard commercial processors.

The implementation of the L2 trigger system varies significantly across experiments, from customised in-house solutions to independent processor farms. The issue encountered by all experiments, which have opted for multiple trigger levels, is the definition of the functionality that the L2 system should provide. Of all the trigger levels after L1, the L2 trigger is the most challenging one, since it has to operate at the highest event rates, often without the benefit of full-granularity and full-resolution data, though with data from more detectors and of higher quality than that used by the L1 Trigger. Decisions that have to be made are the rejection factor that the L2 trigger must provide, the quality of the information it will be provided with, the interconnects between the detector readout, the L1 trigger and the L2 trigger, and finally, the actual implementation of the processing units which will execute the selection algorithms.

Ideally, the High-Level Trigger (HLT) should have no built-in architectural nor design limitations other than the total bandwidth and CPU, which can be purchased based on the experiments resources. Indeed, from very early on, the desire to provide the maximum possible flexibility to the HLT led to the first design principle adopted by ATLAS: the HLT selection should be provided by algorithms executed on standard commercial processors, avoiding all questions and uncertainties related to home-grown hardware processors.

The architecture is depicted schematically in Fig. 16.16. The implementation of the L2 trigger has the advantage that much less data are required to flow into the event filter farm, which in turn has more time to process incoming events. The L2 farm, on the other hand, has to provide a decision on all the events accepted by the L1 trigger. To reduce the data flow into the L2 farm, only a fraction of the detector information is actually transferred from the readout buffers to the L2 processors. This is the concept of the "Region of Interest" (ROI). In brief, the result of the L1 trigger drives the L2 processing, by indicating the regions of the detector which are involved in scrutinising the physics object (electron, muon, jet,...) identified by the L1 trigger. These regions are small, with a total data size of only a few percent of the total event size, so that the full set of data from these regions can be transferred to the L2 farm. The L2 algorithms employ sequential selection and usually not all the data from the ROI in question have to be read in. This farm has tens of *ms* to provide the L2 decision. The events accepted by L2 are sent to the event filter farm, which

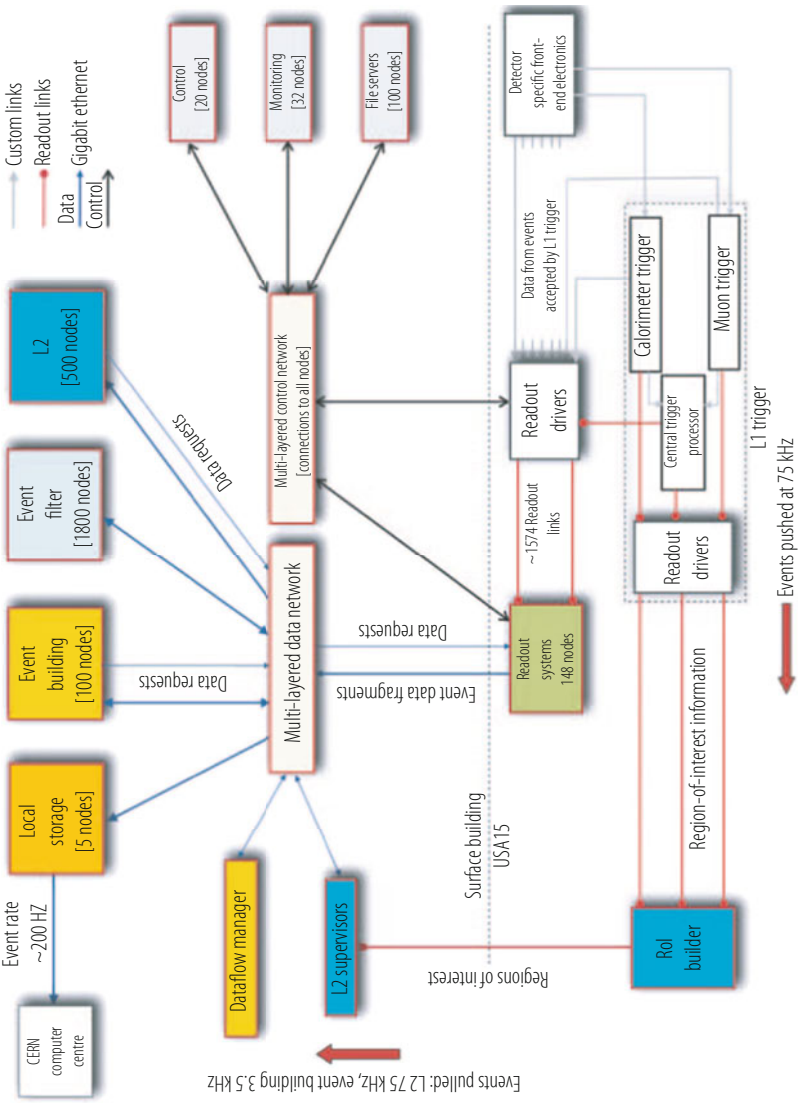


Fig. 16.16 Block diagram of the ATLAS trigger and data acquisition system. Also shown are the different components of the dataflow

now has access to the full event data. This farm runs the final, essentially offline-like selection, "seeding" the reconstruction from the objects previously identified by the L2 trigger in order to reduce the total processing time. The rate input into the event filter farm is a few kHz, so the selection at this level has to provide typically a factor of 10 in rate reduction.

The system relies on commercially available networks for the interconnection between the readout buffers and the HLT farm. The advent of very inexpensive Gbit Ethernet switching fabrics and processor interfaces, along with the rapidly deployable 10 Gbit Ethernet standard, have rendered all early thoughts (back in the mid-1990's) of potential home-grown solutions obsolete.

16.6.2 L1 Trigger System

The L1 trigger has to process information from the detector at the full beam crossing rate of 40 MHz. The very short time between two successive beam crossings (25 ns), along with the wide geographical distribution of the electronic signals from the detector, excludes real-time processing of the full detector data by general-purpose, fully programmable processing elements.

The data are, instead, stored in pipelines awaiting the decision of the L1 trigger within up to 3 μs. The maximum time available for processing in the L1 trigger system is determined by the limited memory resources available in the front-end (FE) electronics which store the detector data during the L1 decision-making process. Technology and financial considerations at the time of the design resulted in a limit of at most 128 bunch crossings, i.e. the equivalent of approximately 3 μs of data, which can be stored in the FE memories. This total latency of 3 μs therefore includes the unavoidable latency components associated with the transfer of the detector information to the processing elements of the L1 trigger and with the latency of the propagation of the L1 decision signals back to the FE electronics. The resulting time available for the actual processing of the data is no more than $\sim 1-1.5\,\mu$s.

In order to avoid dead-time, the trigger electronics must also be pipelined since every process in the trigger must be repeated every 25 ns. The high operational speed and pipelined architecture also imply that only specific data can be brought to the corresponding processing elements in the trigger system. In addition, the data must flow synchronously across the trigger logic in a deterministic manner.

This architecture results in the presence of data from multiple crossings being processed sequentially through the various stages of the trigger logic. To achieve this, most trigger operations are either simple arithmetic operations or functions, which use memory look-up tables, where an address is used to produce rapidly a previously calculated (and stored) result. Moreover, the short time available significantly restricts the data, which can be used in forming the L1 trigger decision, in two ways: on the timing front, the only usable data can come from detectors with very fast response or from slower detectors, which have both good time resolution

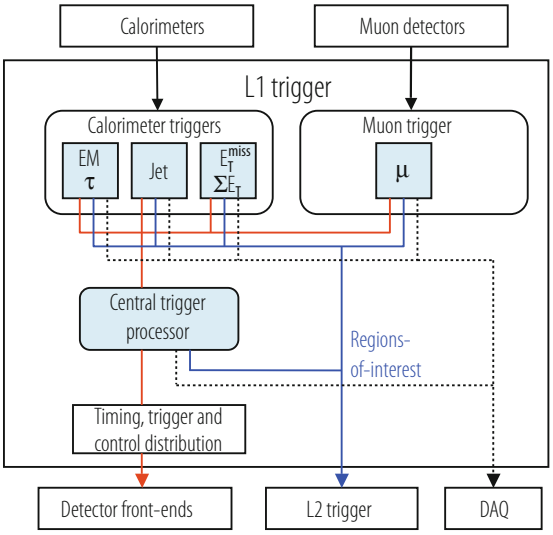


Fig. 16.17 Block diagram of the ATLAS L1 trigger. The overall L1 accept decision is made by the central trigger processor, taking input from calorimeter and muon trigger results. The paths to the detector front-ends, L2 trigger, and data acquisition are shown from left to right in red, blue and black, respectively

and low occupancy; on the volume front, only reduced, coarse information from the calorimeter and muon chambers, corresponding to a smaller fraction of the total volume, and thereby requiring less processing power than e.g. tracker data, can be used.

The block diagram of the ATLAS L1 trigger is shown in Fig. 16.17. It contains a calorimeter trigger, a muon trigger and an overall central trigger processor. The system relies on a Timing, Trigger and Control (TTC) system derived from a precision 40 MHz clock distributed by the LHC accelerator. The different subsystems are essentially independent of each other and the interactions among them are limited to the explicit communication lines in the diagram.

16.6.2.1 Muon Trigger

The L1 muon trigger provides the trigger processor with information on the number, quality and transverse momentum of muon tracks in each event. It consists of a barrel section, two end-cap sections and a part which combines the information from the full system and prepares the input to the central trigger processor. The chambers used in the L1 trigger are used mainly for this purpose, i.e. in the end-cap the L1 muon trigger system uses Thin Gap Chambers (TGC) to cover the region of small angles with respect to the beam axis, whereas, in the barrel, it uses Resistive Plate Chambers (RPC). In both cases, the chambers were selected on their ability to provide signals fast enough for the L1 trigger. Each of the two L1 muon trigger systems has its own trigger logic with different pattern-recognition algorithms.

At the end of processing by the local trigger processors, the muon trigger information from the various sources is collected, and the trigger decision is

prepared before presenting it to the central trigger processor. This intermediate stage carries some significant functionality: the muon trigger to central trigger processor Interface resolves overlaps between chamber sectors in the barrel and between barrel and end-cap chambers and forms the muon candidate multiplicities for the trigger decision.

The final decision on the event is obtained by the central trigger processor itself, using either information from only the muon trigger or in association with other objects in the event (e.g. the presence of a high-p_T electron).

16.6.2.2 Calorimeter Trigger

The L1 calorimeter trigger provides essentially all the L1 trigger streams for the experiment (electrons, photons, QCD jets, $\tau-$jets, missing E_T) except for the muons. The architecture of this trigger contains three elements, namely the generation of the trigger primitives, a local calorimeter trigger which processes information from limited parts of the detector, and a global calorimeter trigger which combines all the information from the local processors, prior to sending the summaries to the central trigger processor. Data from the calorimeters are combined to form trigger towers of approximate size 0.1×0.1 in $\eta - \phi$ space. Analogue sums are formed on the detector and sent through analogue transmission to the counting room.

The information is then digitised and processed to determine the transverse energy E_T in each trigger tower. As discussed previously, most of the ATLAS calorimeters have pulse shapes which extend well beyond a single crossing, so the signals are processed to assign each energy deposition to the correct bunch crossing. Once the transverse energies and the bunch crossing are determined, the algorithms in the local calorimeter trigger take over. The basic features can be summarised as follows:

- Electrons and photons are searched for as peaks in the E_T deposited in a limited $\eta - \phi$ region (neighboring towers) of the EM calorimeter. The corresponding hadronic energy is required to be small, relatively to the EM calorimeter energy. Additional isolation requirements, e.g. by demanding that neighbouring towers do not have energy larger than a certain threshold, may be imposed.
- Jets are formed by adding the energy in a large $\eta - \phi$ region consisting of an array of 4×4 trigger towers/elements. The algorithm provides flexibility in the measurement of the jet energy through the use of a sliding window, but therefore requires an additional processing step to settle jet overlaps and eliminate double-counting.
- τ-jets are formed by demanding very narrow energy depositions in the electromagnetic and hadronic calorimeters. Isolation requirements may also be applied.
- The missing transverse energy (as well as the total transverse energy in the event) is estimated from the sum of the transverse energies of all the calorimeter cells.

The sum of the transverse energies of all jets found in the event is also provided; this will be more stable with increasing luminosity than the sum over all cells.

The results of this local processing, i.e. the electron/photon, τ-jet, and jet candidates are passed on to the central trigger processor. The physics objects are sorted in E_T and finally used in the global decision, possibly in association with other L1 objects in the event.

16.6.3 High-Level Trigger and Data Acquisition Systems

Experience with the data acquisition (DAQ) systems of previous experiments at high-energy lepton and hadron colliders has resulted in the establishment of several fundamental design principles which have been embedded in the architecture from the very beginning.

The technological advances witnessed over the last 20 years have progressed at an extraordinary rate, which until now has remained constant with time. It was decided to invest in these advances of technology and especially in the two main fronts that drive them, processing power and network speed. An additional consideration has been the expected evolution of the experiment and its data acquisition system, rendering a fully programmable HLT system highly desirable to avoid major design changes. The added flexibility provided by the fully programmable environment of a standard CPU also implies that algorithmic changes necessary for the resolution of unforeseen backgrounds or other adverse experimental conditions can be easily introduced. A final consideration was the desire to minimise the amount of non-standard, in-house solutions.

As a result of the above considerations, the data acquisition system relies on industrial standards to the greatest possible extent, and employs commercially available components, if not full-fledged systems, wherever these could meet the requirements. This applies to both hardware and software systems. The benefits of this decision are numerous, with the most important ones being the resulting economies in both the development and production costs, the prompt availability of the relevant components from multiple competing sources, and a maintenance and support mechanism which does not employ significant in-house resources.

Another general design principle, adopted at the very earliest stages of development, is that of maximal scaling. This addresses the fact that the accelerator conditions, the experimental conditions, and finally the physics programme itself are all expected to evolve significantly with time. An easily scalable system is one in which the functions, and thus the challenges as well, are factorised into sub-systems with a performance independent of the rest of the system.

The long difference in time between the design of the systems and their final implementation and deployment implied a development cycle different from that of the other detector projects. In the case of the DAQ systems, the understanding of the required functionality of the various elements of the system was, in many

cases, separated from their performance. The numerous and challenging subsystem components were thus developed along two independent paths. The first development path concentrated on the identification and implementation of the full functionality needed for operation in the final DAQ. The second path concentrated on the issues that arise when the functions identified in the first path are executed at the performance levels required by the final DAQ system.

Following these principles, ATLAS has pursued an R&D programme, which has resulted in a system that could be implemented for the early luminosities of the LHC, and could be scaled to the expected needs at the full design luminosity, since the system architecture is such that in a number of incremental steps, the performance of the system can be increased proportionally.

16.6.3.1 Data Acquisition

The main elements of the ATLAS DAQ system are described in more detail below:

- Detector readout system: this consists of modules which read the data corresponding to a single bunch crossing out of the front-end electronics upon the reception of a L1 trigger accept signal. There are approximately 1600 such modules in the ATLAS readout.
- Event builder: this is the collection of networks, which provide the interconnections between the detector readout and the HLT. It provides (and monitors) the data flow and employs a large switching fabric. ATLAS has two such networks, one for the L2 trigger and one for the event filter.
- HLT systems: these are the processors, which deal with the events provided by the detector readout. They execute the HLT algorithms to select the events to be kept for storage and offline processing.
- Controls and monitors: these consist of all the elements needed to control the flow of data (events) through the DAQ system, as well as the elements needed to configure and operate the DAQ. This includes all the provisions for special runs, e.g. for calibrations, that involve special setups for both the detectors, the trigger and the readout. The other major functionality is the monitoring of the various detector elements, of the operation of the L1 and HLT and of the state of the DAQ system and its elements.

The factorisation of the DAQ function into tasks, which can be made almost independent of each other, facilitates the design of a modular system which can be developed, tested and installed in parallel. To ensure this factorisation, the different operational environments of the four functional stages must be decoupled. This is achieved via the introduction of buffering of adequate depth in between each of these stages. The primary purpose of these buffers is to match the very different operating rates of the elements at each stage. As an example, at a rate of 100,000 events per second, the readout system delivers an event every $10\,\mu s$. On the other hand, the event building process requires, even assuming a 100% efficiency of 2 Gb/s links,

a time of ∼ms to completely read in the event. This is therefore the rate at which the elements of the farm system can operate on events. The two time-scales are very different, and this is where the deep buffers present in the readout system serve to minimise the coupling between the stages.

The design of the DAQ system is very modular, thereby allowing for a staged installation. The event builder has been conceived with the possibility of a phased installation from the very beginning. The operation of the ATLAS experiment has begun with a DAQ system serving only a reduced bandwidth of approximately 20–40 kHz. The deferrals were necessary because of funding pressures, whereas a staged installation of the DAQ was viewed as less damaging to the physics programme, since the initial instantaneous luminosity of the LHC is far below the design value.

16.6.3.2 High-Level Trigger

As mentioned previously, the HLT is a software filtering process executed on standard commercially available processors. The software is drawn from the offline reconstruction software of the experiment. Both levels of the HLT are executed within the offline framework, but in contrast to the event filter which uses the same algorithms as the offline, the L2 trigger processors run more dedicated code (in particular with faster data-preparation algorithms). The trigger software is steered differently from the offline and initiates the reconstruction from the physics candidate objects identified by the previous levels (L1 or L2 trigger). The overall rejection factor is achieved by applying, in software, a number of successive reconstruction and selection steps.

As an example, the HLT electron trigger is typically driven by a L1 electron/photon candidate, which is identified as a high-energy isolated electromagnetic (EM) energy deposition in the calorimeters. At the output of the L1 trigger, the rate is dominated by QCD jets. The first task in reconstructing the electron in the HLT is to rerun the clustering algorithm with access to the full granularity and resolution of the EM calorimeter and to obtain a new, more accurate, measurement of the transverse energy (E_T) of the EM cluster. Given the rapidly falling cross section, this already provides a rejection factor of ≈2 with respect to the input event rate. Further shower-shape and isolation cuts are also applied at this point. The events surviving the EM calorimeter requirements are subsequently subjected to a search for a charged-particle track in the tracking detectors. The matching between track and cluster is a powerful requirement, which yields at least a factor of 10 rejection against jets while maintaining a very high efficiency.

Events selected by the HLT are forwarded to mass storage and from there to the offline system for reconstruction and physics analysis. Given the unprecedented rate of online rejection, another very important task of the HLT is to provide detailed information on the events which have been rejected at each stage of the filtering process.

16.7 Computing and Software

The ATLAS computing and software infrastructure is clearly of paramount importance. The functionality and flexibility of both will determine, to a very large extent, the rate and quality of the physics output of the experiment. As expected, there are numerous challenges to be addressed also in these two areas.

On the computing side, the LHC experiments represent a new frontier in high-energy physics. What is genuinely new at the LHC is that the required level of computing resources can only be provided by a number of computing centres working in unison with the CERN on-site computing facilities. Off-site facilities will thus be vital to ATLAS operation to an extent that is completely different from previous large experiments. Usage of these off-site facilities necessitates the substantial use of Grid computing concepts and technologies [33]. The latter allow for the sharing of the responsibility for processing and storing the data, but also for providing the same level of data access, and making available the same amount of computing resources to all members of the collaboration.

A second challenge for computing is the development and operation of a data storage and management infrastructure which is able to meet the demands of a yearly data volume of O(10) Petabytes and is used by both organised data processing and individual analysis activities, which are geographically dispersed around the world.

The architecture which is now in place is geographically distributed and relies on four levels or tiers, as illustrated in Fig. 16.18. Primary event processing occurs at CERN in the so-called Tier-0 facility. Raw data are archived at CERN and sent (along with the reconstructed data) to the Tier-1 centres around the world. These centres share among themselves the archiving of a second copy of the raw data, while they also provide the reprocessing capacity and access to the various versions of the reconstructed data, and allow scheduled analysis of the latter by physics analysis groups. A more numerous set of Tier-2 centres, which are smaller but still have substantial CPU and disk storage resources, provide capacity for analysis, calibration activities and Monte Carlo simulation. Datasets, which are produced at the Tier-1 centres by physics groups, are copied to the Tier-2 facilities for further analysis. Tier-2 centres rely upon the Tier-1 centres for access to large datasets and secure storage of the new data they produce. A final level in the hierarchy is provided by individual group clusters used for analysis: these are the Tier-3 centres.

The ATLAS collaboration also relies on the CERN Analysis Facility (CAF) for algorithmic development work and a number of short-latency data-intensive calibration and alignment tasks. This facility is also expected to provide additional analysis capacity with, as an example, re-processing of the express-stream data and short turn-around analysis jobs.

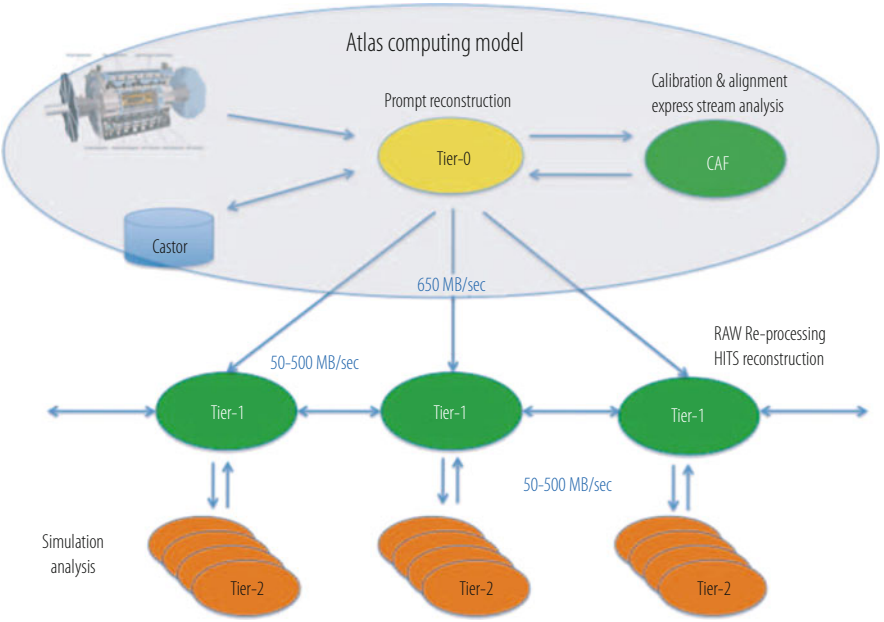


Fig. 16.18 Schematic flow of event data in the ATLAS computing model, illustrating the Tier-0, Tier-1 and Tier-2 connections. Tier-3 centres (typically smaller analysis clusters) are not included

16.7.1 Computing Model

The tasks of archiving, processing and distributing the ATLAS data across a worldwide computing organisation are of an unprecedented magnitude and complexity. The ever-present financial limitations, along with the unpredictability of the accelerator and detector operational details at the start-up, have implied the creation of a very flexible yet cost-effective plan to manage all the computing resources and activities. This plan, referred to as the computing model, was difficult to set up initially since the resources for computing had not been included in the initial funding plan for the LHC experiments. Over the past 5 years, however, a detailed computing model has been put in place and tested thoroughly with large-scale samples of simulated data and various technical computing challenges. This computing model describes as accurately as feasible the flow of data from the data acquisition system of the experiment to the individual physicist desktop [30]. Over the past few years, it has adapted to the evolution of the major parameters which govern it, such as the respective sizes of the various data types, the reality of the resources available at the various Tiers, and the more and more precise understanding of the requirements of the actual analysis in the various physics domains.

The main requirement on the computing model is to provide prompt access to all the data needed to carry out physics analyses. This typically translates to providing all members of the collaboration with access to reconstructed data and appropriate, more limited, access to raw data for organised monitoring, calibration and alignment activities. As already mentioned, the key issue is the decentralisation and wide geographic distribution of the computing resources. Sharing of these resources is possible through the Grid and its middleware, and therefore the interplay with the Grid is built into the models from the very beginning.

The most important elements of the computing model are the event data model and the flow of the various data types to the analysis processes.

16.7.1.1 Event Data Model

The physics event store contains a number of different representations, or levels of detail, of the physics events from the raw (or simulated) data all the way to reconstructed and summary data suitable for massive fast analysis. The different types of data are:

- Raw data: this is the byte-stream output of the High-Level Trigger (HLT) and is the primary input to the reconstruction process. The ATLAS experiment expects ≈1.5 MB of data arriving at a rate of ≈200–300 Hz. Events are transferred from the HLT farm to the Tier-0 in 2 GB files containing events from a data-taking period with the same trigger selections from a single LHC fill. The events will generally not appear in a consecutive order, since they will have undergone parallel processing in the HLT farm beforehand.
- Reconstructed data (referred to as Event Summary Data or ESD): this is the output of the reconstruction process. Most detector and physics studies, with the exception of calibration and alignment procedures, will only have to rely on this format. The data are stored using an object-oriented (OO) representation in so-called POOL-format files [31, 32]. The target size for the ESD files has increased from 500 to 800 kB per event over the past few years.
- Analysis Object Data or AOD: this is derived from the ESD format and is a reduced event representation, intended to be sufficient for most physics analyses. The target size is roughly a factor five smaller than that of the ESD (i.e. 100–200 kB per event) and the contents are physics objects and other high-level analysis elements.

If experience from the Tevatron and initial experience from the experiment commissioning and early data-taking phase are used as a guide, it is expected that in the early stages of the machine and experiment commissioning the ESD format will be in heavy use. The AOD format is expected to become the dominant tool for studies only when both machine and experiments are in steady-state data-taking. Nevertheless, it is planned to commission the AOD format with real collision data as early as possible, since one of the biggest constraints on the computing model will be the access bandwidth to the data. The AOD, in addition to being the format

with the smallest size, has, by construction, the most compact and complete physics information of the event, and is thus going to be indispensable in carrying out high-statistics analyses.

In preparation for the hopefully soon-to-come high-statistics analysis era, ATLAS has defined two further formats, namely a condensed data format for tagging events with certain properties, called TAG, and a Derived Physics Data format(or DPD), which are intended for use in end-user analyses. TAG data are event-level metadata, i.e. thumbnail information about each event to enable rapid and efficient selection for individual analyses. The TAG data are also stored in a relational database to enable various searches via database queries. The average size is a few kB per event. The DPD format corresponds to the highest-level of data representation, with "ntuple"-like content, for direct analysis and display by analysis programs.

These official data formats have been deployed as the vehicle for running physics analyses. As an example, the AOD format and its contents have been the subject of several generations of very extensive sets of tests with different data, conditions, and subsequent uses. Of course, since the AOD format contains only a subset of the information in the event, there will always be analyses that need to refer back to the ESD format. The most critical part of the optimisation of these various formats over the past few years has therefore been to select appropriately the objects to be included in the AOD. There is usually a trade-off between storage cost and CPU to derive the additional objects to be studied, and the details depend very strongly on the sample size required and the number of times the sample is used.

16.7.1.2 Data Flow and Processing

To maximise the physics reach of the experiment, the HLT farms will write events at the maximum possible data rate, which can be supported by the computing resources. Currently, this is expected to be in the range of 200–300 Hz, essentially independent of the instantaneous luminosity of the accelerator. Trigger thresholds will be adjusted up or down to match the maximum data rate, in order to maintain consistency with the data storage and processing capabilities of the offline systems. Extensive test campaigns have shown that the online-offline link and the Tier-0 centre are able to keep up in real-time with the HLT output rate.

The HLT output is streamed according to trigger type for the subsequent reconstruction and physics analysis. In addition, specialised calibration streams allow for independent processing from the bulk of the physics data. These streams are required to produce calibration and alignment constants of sufficient quality to allow a useful first-pass processing of the physics streams with minimum latency. ATLAS also makes use of an express stream, which is a set of physics triggers corresponding to about 5% of the full data rate. These events are selected to tune the physics and detector algorithms and also to provide rapid updates to the calibration and alignment constants required for the first-pass processing.

Streams can be used for a variety of purposes. The primary use, as mentioned previously, is to allow the prioritisation of the processing of the data. As an example, having the di-muon dataset as a independent stream obviously results in a much faster turnaround on any analysis that relies on these data. Streams can also be useful in the commissioning phase, to debug both the software and the overall online and offline computing systems. As an example, a special "debug" stream is dedicated to problematic events, e.g. failing in the HLT step, to facilitate the understanding of errors in the system. Obviously, such streams will be created as the need arises, will be rate-limited, and may even be withdrawn once the primary motivation for them is no longer present.

The first step before full-fledged prompt reconstruction is the actual processing of the calibration data in the shortest possible time. The plan calls for a short 1 to 2-day latency in completing this task. Once the calibration and alignment constants are in place, a first-pass (or prompt) reconstruction is run on the primary event streams, and the resulting reconstructed data (ESD and AOD formats) are archived into the CERN mass storage system.

Upon completion of this step, the data are distributed to the Tier-1 centres. Each Tier-1 site assumes responsibility for a fraction of the reconstructed data. Most of the ESD format data are, however, not available on disk for individual user access. A major role for the Tier-1 centres is the reprocessing of the data, once more mature calibrations and software are available, typically once or twice every year. By shifting the burden of reprocessing to the Tier-1 centres, the experiment can reprocess its data asynchronously and concurrently with data-taking and the associated prompt processing. The Tier-2 centres can obtain partial or full copies of the AOD/DPD/TAG format data, which will be the primary tool for physics analysis. The Tier-2 centres will also be responsible for large-scale simulation tasks, once the Tier-1 sites will be very busy with data reprocessing.

16.7.2 Software

On the software front, there have been two major issues encountered by the LHC experiments, which are either new or simply appear to a much greater extent than in the past: the distributed nature of the development and the maintainability of the code over long time-scales:

- Software development has had to continue down the path established at LEP and at the Tevatron: the code is developed in a distributed manner with responsibilities that span multiple individuals, institutions, countries and time zones. While for the large-scale hardware projects, a factorisation of the overall construction into substantial units has been possible, software, with its much wider contributor base within the collaborations, has a larger degree of fragmentation. This has necessitated the formation of intricate project structures to monitor and steer

the code development. The usual issues which result from relying on multiple institutions and funding agencies have risen here as well (see Sect. 16.2).
- Another major issue has been the maintainability of the systems. Given the expected long lifetime of the LHC programme, it was deemed necessary, from the very beginning, that the software systems be built using object-oriented methodologies. The C++ programming language has been chosen as the major development tool.

At the heart of the software system of the experiment is the software framework, which provides support for all the data-processing tasks. All such tasks, including the simulation, reconstruction, analysis, visualisation, and, very importantly, the high-level trigger operate within this framework. It provides the basic software environment in which code is developed and run, as well as all the basic services (e.g. access to calibration and conditions data, input/output facilities, persistency, to name but a few examples).

All the applications, which are built on top of the framework use a component model, i.e. they have building blocks, which appear to the framework as standard plug-ins. The main advantage of the component model is the factorisation of any one solution into a number of independent software codes, but also a significant flexibility to adapt to changes in the future. The final major architectural and design principle has been the separation of algorithms from the data and the acceptance of different data representations in memory (transient) and file storage (persistent).

16.7.3 Analysis Model

As has been already mentioned, the ESD and AOD/DPD formats are the primary tools for carrying out physics studies. Both formats are stored in POOL files and are processed using the respective software framework of each experiment. The decreasing event size in the event model allows the users to process a much larger number of AOD/DPD events than ESD events. In addition, the AOD/DPD formats will be more accessible, with a full copy at each Tier-1 site and large samples at Tier-2 sites. It is therefore expected that most analyses will be carried out on AOD/DPD data.

To illustrate the ATLAS analysis model with a concrete example, a specific analysis task may begin with a query against the TAG data to select a subset of events for processing using a suitable DPD format. This query might be for events with two leptons, missing transverse energy and at least two jets, all above certain thresholds. The result of this query is then used to define a dataset (or set of files) containing the information for these events. The analysis would then proceed to make further event selection by refining various physics quantities, e.g. the muon isolation or the missing transverse energy calculation. The fine-grained details of how much processing and event selection will be carried out by individuals versus organised physics groups (e.g. the Higgs group) is not frozen yet. It is widely expected that

both modes of operation will occur, i.e. that there will be data samples, which are selected and perhaps processed further in an organised manner by large groups of the collaboration, but also samples created by individuals. The relative fraction of each will be driven to a large extent by the resources that will be available at any given time.

The last element of the analysis model is a distributed analysis system which allows for the remote submission of jobs from any location. This system splits, in an automated way, an analysis job into a number of smaller jobs that run on subsets of the input data. The results of the job may be merged to form an output dataset. Partial results from these jobs are made available to the user before the full set of jobs runs to completion. Finally, the distributed analysis system will ensure that all jobs and resulting datasets are properly catalogued for future reference.

16.8 Expected Performance of Installed Detectors

16.8.1 Tracker Performance

Table 16.12 shows a comparison of the main performance parameters of the ATLAS and CMS trackers, as obtained from extensive simulation studies performed over the years and bench-marked using detailed test-beam measurements of production modules wherever possible. The unprecedentedly large amount of material present

Table 16.12 Main performance characteristics of the ATLAS and CMS trackers

	ATLAS	CMS
Reconstruction efficiency for muons with $p_T = 1\,\text{GeV}$	96.8%	97.0%
Reconstruction efficiency for pions with $p_T = 1\,\text{GeV}$	84.0%	80.0%
Reconstruction efficiency for electrons with $p_T = 5\,\text{GeV}$	90.0%	85.0%
Momentum resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 0$	1.3%	0.7%
Momentum resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 2.5$	2.0%	2.0%
Momentum resolution at $p_T = 100\,\text{GeV}$ and $\eta \approx 0$	3.8%	1.5%
Momentum resolution at $p_T = 100\,\text{GeV}$ and $\eta \approx 2.5$	11%	7%
Transverse i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 0$ [μm]	75	90
Transverse i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	200	220
Transverse i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 0$ [μm]	11	9
Transverse i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	11	11
Longitudinal i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 0$ [μm]	150	125
Longitudinal i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	900	1060
Longitudinal i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 0$ [μm]	90	22–42
Longitudinal i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	190	70

Examples of typical reconstruction efficiencies, momentum resolutions and transverse and longitudinal impact parameter (i.p.) resolutions are given for various particle types, transverse momenta and pseudorapidities

in the trackers is reflected in the overall reconstruction efficiency for charged pions of low transverse momentum, which is only slightly above 80%, to be compared to 97% obtained for muons of the same transverse momentum. The electron track reconstruction efficiency is even more affected by the tracker material and the numbers shown in Table 16.12 for electrons of 5 GeV transverse momentum are only indicative, since the efficiency obtained depends strongly on the criteria used to define a reasonably well measured electron track. The somewhat lower efficiencies obtained in the case of CMS are probably due to the higher magnetic field, which enhances effects due to interactions in the detector material. The combined performance of the tracker and electromagnetic calorimeter is discussed in Sect. 16.8.2.

The higher and more uniform magnetic field and the better measurement accuracy at large radius of the CMS tracker result in a momentum resolution on single tracks, which is better than that of ATLAS by a factor of almost 3 over the full kinematic range of the fiducial acceptance of the trackers. The impact parameter resolution in the transverse plane is expected to be similar at high momenta for both trackers, because the smaller pixel size in ATLAS is counter-balanced by the charge-sharing between adjacent pixels and the analogue readout in the CMS pixel system. In contrast, the smaller pixel size of the CMS tracker in the longitudinal dimension leads to a significantly better impact parameter resolution in this direction at high momenta.

In summary, the ATLAS and CMS trackers are expected to deliver the performances expected at the time of their design, despite the very harsh environment in which they will operate for many years and the difficulty of the many technical challenges encountered along the way. In contrast to most of the other systems, however, they will not survive nor deliver the required performance if the LHC luminosity is upgraded to $10^{35}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The ATLAS and CMS trackers will therefore have to be replaced by detectors with finer granularity to meet the challenges of the higher luminosity and with an order of magnitude higher resistance to radiation. This will be the major upgrade challenge for both experiments and a lively programme of research and development work has already been launched to this end.

16.8.2 Calorimeter Performance

The performance to be expected in situ for the very large-scale calorimeter systems of ATLAS and CMS is difficult to directly extrapolate from test-beam data. The calibration of these complex electromagnetic and hadronic calorimeter systems can indeed be to some extent ported with high precision from the test-beam measurements to the actual experiment and, more importantly, performed in situ using a set of benchmark physics processes such as $Z \rightarrow ee$ decays and $W \rightarrow jet - jet$ decays. This situation is somewhat new because of the following reasons:

- for the first time, there will be the possibility to control the absolute scale of hadronic jet energy measurements by using sufficiently abundant statistics from $W \to jet - jet$ topologies occurring in top-quark decays.
- extensive test-beam measurements in configurations close to that of the real experiment will have been performed at the time of first data-taking.
- it should be possible to constrain the absolute scale of the overall hadronic calorimetry using the measured response to charged pions of energies between 1 and 300 GeV and controlling this scale in situ, using a variety of samples, from single isolated tracks at the lower end of the range to e.g. clean samples of $\tau \to \pi^{\pm}\nu$ decays.

During the past 15 years, a large-scale and steady software effort has been maintained in the collaborations to simulate in detail calorimeters of this type well before they begin their operation. The complex geometries and high granularities described above and the high energies of the products of the collisions have naturally augmented considerably the computing effort required to produce large-statistics samples of fully simulated events. A few examples are shown below for photon, electron, jet and missing transverse energy measurements.

16.8.2.1 Electromagnetic Calorimetry

Figure 16.19 shows an example of the expected precision with which photon energy measurements will be performed in ATLAS (left) and CMS (right) over the energy range of interest for $H \to \gamma\gamma$ decays. In the case of ATLAS, the results are shown for all photons (unconverted and converted) and for three values of pseudorapidity. In the case of CMS, the results are shown for dominantly unconverted photons in the barrel crystal calorimeter. The selected photons are required in this latter case to have deposited more than 94.3% of their energy in a 3 by 3 crystal matrix normalised to the 5 by 5 crystal matrix used to compute the total energy. This basically selects unconverted photons and some late conversions with a 70% overall efficiency. For a photon energy of 100 GeV, the ATLAS energy resolution varies between 1.0 and 1.4%, depending on η. These numbers increase respectively to 1.2 and 1.6% if one includes the global constant term of 0.7%. The overall expected CMS energy resolution in the barrel crystal calorimeter is 0.75% for the well-measured photons at that energy (Fig. 16.19 includes the global constant term of 0.5%). This example shows that the intrinsic resolution of the CMS crystal calorimeter is harder to obtain with the large amount of tracker material in front of the EM calorimeter and in the 4T magnetic field: between 20 and 60% of photons in the barrel calorimeter acceptance convert before reaching the front face of the crystals.

Similarly, Fig. 16.20 shows an example of the expected precision with which electron energy measurements will be performed in ATLAS (left) and CMS (right). In the case of ATLAS, the results are shown for electrons at $\eta = 0.3$ and 1.1 in the energy range from 10 to 1700 GeV. The energy of the electrons is always collected in a 3 by 7 cell matrix, which, as for the photons, is wider in the bending direction

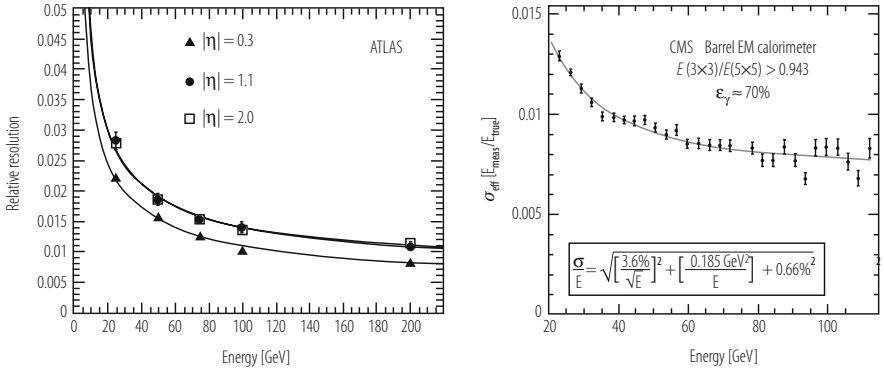


Fig. 16.19 For ATLAS (left) and CMS (right), expected relative precision on the measurement of the energy of photons reconstructed in different pseudorapidity regions as a function of their energy (see text). Also shown are fits to the stochastic, noise and local constant terms of the calorimeter resolution

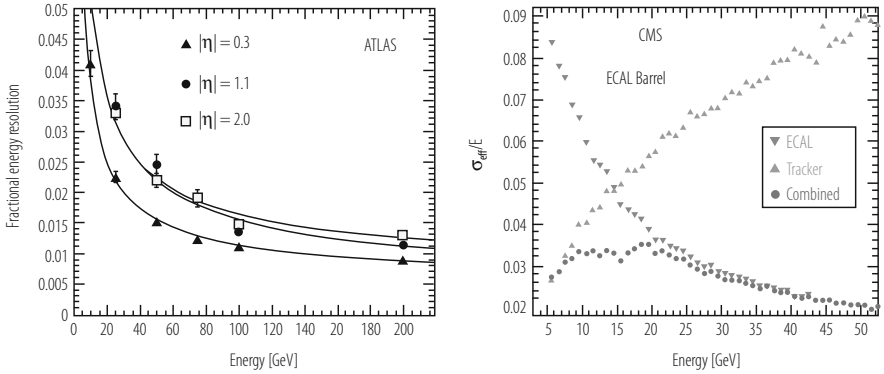


Fig. 16.20 For ATLAS (left) and CMS (right), expected relative precision on the measurement of the energy of electrons as a function of their energy over the energy range of interest for $H \rightarrow ZZ^{(*)} \rightarrow eeee$ decays. In the case of ATLAS, the resolution is shown for three values of pseudorapidity (only the electron energy measurement is used, with the energy collected in a 3 by 7 cell matrix in $\eta \times \phi$ space), together with fits to the stochastic and local constant terms of the calorimeter resolution. In the case of CMS, the combined (tracker and EM calorimeter) effective resolution at low energy, taken as the r.m.s. spread of the reconstructed energy, collected in a 5 by 5 cell matrix and normalised to the true energy, is shown over the acceptance of the barrel crystal calorimeter, together with the individual contributions from the tracker and the EM calorimeter

to collect as efficiently as possible the bremsstrahlung photons while preserving the linearity and low sensitivity to pile-up and noise. In the case of CMS, the effective resolution (r.m.s. spread) is shown for the barrel crystal calorimeter and in the most difficult low-energy range from 5 to 50 GeV. Refined algorithms are used, in both the tracker and the calorimeter, to recover as much as possible the bremsstrahlung tails and thereby to restore most of the excellent intrinsic resolution of the crystal

calorimeter. Nevertheless, for electrons of 50 GeV in the barrel region, the ATLAS energy resolution varies between 1.3% (at $\eta = 0.3$) and 1.8% (at $\eta = 1.1$) without any specific requirements on the performance of the tracker at the moment. In contrast, the CMS effective resolution is estimated to be 2%, demonstrating that it is harder to reconstruct electrons, with a performance in terms of efficiency and energy resolution similar to that obtained in test beam, than photons.

Further performance figures of critical importance to the electromagnetic calorimeters are those related to electron and photon identification in the context of overwhelming backgrounds from QCD jets and of pile-up at the LHC design luminosity, of γ/π^0 separation, of efficient reconstruction of photon conversions and of measurements of the photon direction using the calorimeter alone wherever the longitudinal segmentation provides a sufficiently accurate measurement. All these aspects rely heavily on the details of the longitudinal and lateral segmentation of the EM calorimetry and the reader is referred to the ATLAS and CMS detector performance reports [13, 27] for more information.

Another important issue, especially for the EM calorimeters is the calibration in situ, which will eventually provide the final calibration constants required e.g. for searches for narrow states, such as $H \rightarrow \gamma\gamma$ decays. These can be divided into an overall constant defining the absolute scale and a set of inter-calibration constants between modules or cells:

- the ATLAS EM calorimeter has been shown to be uniform by construction to about 0.4% in areas of 0.2×0.4 or larger in $\Delta\eta \times \Delta\phi$ space. One will therefore have to calibrate in situ only about 440 sectors of this size. The use of the Z mass constraint alone without reference to the tracking should be sufficient to achieve an inter-calibration to better than 0.3% over a few days at low luminosity. If additional problems arise because of the material in the tracker, the use of electrons from W decay to measure E/p will provide additional constraints.
- the CMS crystals could not be pre-calibrated in the laboratory with radioactive sources to better than 4.5%. This inter-calibration spread has been brought down to significantly smaller values using cosmic rays. Without an individual calibration of the crystals in the test beam, one has to rely on in situ calibration for further improvements. Using initially large samples of minimum bias events (including explicit reconstruction of π^0 and η decays) and low E_T jets at fixed η, the inter-calibration could be improved to 1.5% within ϕ-rings of 360 crystals. At a later stage, high statistics samples of W-boson decays to electrons will be needed to reach the target constant term of 0.5%.
- a key issue for both ATLAS and CMS will be to keep the constant term below the respective target values of 0.7 and 0.5% in the presence of the unprecedented amount of material in the trackers. For ATLAS, other major potential contributions to the constant term (each one of the order of 0.2 to 0.3%) are mostly short-range (detector geometry, such as ϕ-modulations, variations of the sampling fraction in the end-caps, absorber and gap thickness fluctuations, fluctuations in the calibration chain, differences between calibration and physics signal), but the more potentially worrisome one is long-range and is related

to the signal dependence on temperature. The LAr signal has a temperature dependence of −2% per degree: the temperature monitoring system in the barrel sensitive volume should therefore track temperature changes above ±0.15°, which is the expected dispersion from the heat influx of 2.5 kW per cryostat. In CMS, the temperature control requirements are even more demanding, since the temperature dependence of a crystal and its readout is about −4.3% per degree for a heat load of 2 W per channel or 160 kW total. The very sophisticated cooling scheme implemented in the super-modules has demonstrated the ability to maintain the temperature to better than ±0.05° and thereby to meet these stringent requirements. Time-dependent effects related to radiation damage of the CMS crystals will have to be monitored continuously with a stable and precise laser system.

16.8.2.2 Hadronic Calorimetry

The expected performance for reconstructing hadronic jets is shown in Fig. 16.21. In the case of ATLAS, the jet energy resolution is depicted for two different pseudorapidity bins over an energy range from 15 to 1000 GeV for two different sizes of the cone algorithm used. The jet energies are computed using a global weighting technique inspired by the work done in the H1 collaboration [28]. In the case of CMS, the jet energy resolution is shown as a function of the jet transverse energy, for a cone size $\Delta R = 0.5$ and for $|\eta| < 1.4$, over a transverse energy range from 15 to 800 GeV. For hadronic jets of typically 100 GeV transverse energy, characteristic for example of jets from W-boson decays produced through top-quark decay, the ATLAS energy resolution varies between 7 and 8%, whereas

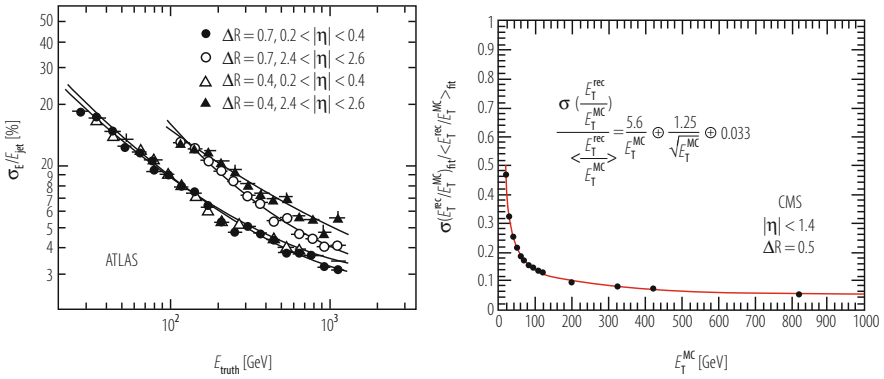


Fig. 16.21 For ATLAS (left) and CMS (right), expected relative precision on the measurement of the energy of QCD jets reconstructed in different pseudorapidity regions as a function of E_{truth}, where E_{truth} is the true jet energy, for ATLAS, and of E_T^{MC}, where E_T^{MC} is the true jet transverse energy, for CMS (see text)

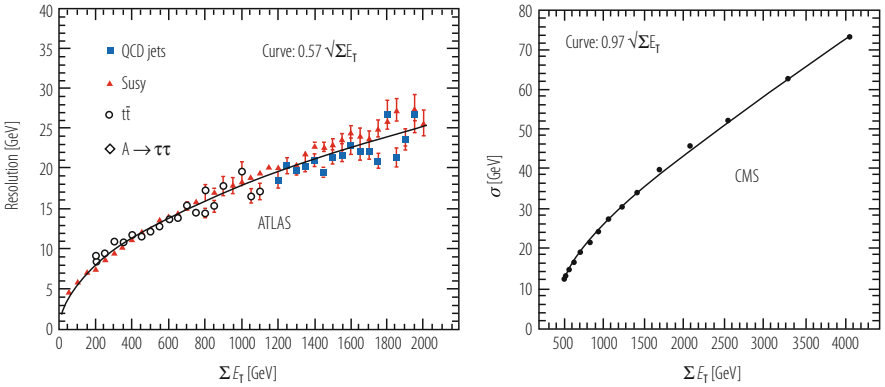


Fig. 16.22 For ATLAS (left) and CMS (right), expected precision on the measurement of the missing transverse energy as a function of the total transverse energy, ΣE_T, measured in the event (see text)

the CMS energy resolution is approximately 14%. The intrinsic performance of the CMS hadron calorimeter can be improved using charge particle momentum measurements, a technique often referred to as particle flow, which was developed at LEP [23]. Initial studies indicate that the jet energy resolution can be significantly improved at low energies, typically from 17 to 12% for $E_T = 50\,\text{GeV}$ and $|\eta| < 0.3$, but such large improvements are not expected for jet transverse energies above 100 GeV or so.

Finally, Fig. 16.22 illustrates a very important aspect of the overall calorimeter performance, namely the expected precision with which the missing transverse energy in the event can be measured in each experiment as a function of the total transverse energy deposited in the calorimeter. The results for ATLAS are expressed as the σ from Gaussian fits to the (x,y) components of the E_T^{miss} vector for events from high-p_T jet production and also from other possible sources containing several high-p_T jets. In the case of CMS, where the distributions are non-Gaussian, the results are expressed as the r.m.s. of the same distributions for events from high-p_T jet production. For transverse momenta of the hard-scattering process ranging from 70 to 700 GeV, the reconstructed ΣE_T ranges from about 500 GeV to about 2 TeV. The difference in performance between ATLAS and CMS is a direct consequence of the difference in performance expected for the jet energy resolution.

16.8.3 Muon Performance

The expected performance of the muon systems has been a subject of very intense study in both experiments. Simulations which take into account a huge amount of

detail from the real geometries of all the chambers and support structures have been refined repeatedly over the years.

In ATLAS, the quality of the stand-alone muon measurement relies on detailed knowledge of the material distribution in the muon spectrometer, especially for intermediate-momentum muons. Reconstruction of these with high accuracy and without introducing a high rate for fake tracks, has to take into account multiple scattering of the muons and thus the details of the material distribution in the spectrometer. This necessitates a very detailed mapping of the detector and the storage of this map for use by the offline simulation and reconstruction programs. The corresponding effect in CMS is much smaller, since the amount of iron in between the muon stations dominates by far and the details of the material are necessary only in the boundaries between the iron blocks.

Figures 16.23 and 16.24 show the expected resolution on the muon momentum measurement. The expected near-independence of the resolution from the pseudorapidity in ATLAS, along with the degradation of the resolution at higher η in CMS are clearly visible. The resolution of the combined measurement in the barrel region is slightly better in CMS due to the higher resolution of the measurement in the tracking system, whereas the reverse is true in the end-cap region due to the better coverage of the ATLAS toroidal system at large rapidities. A summary of the performance of the two muon measurements can be found in Table 16.13 for muon momenta between 10 and 1000 GeV.

The expected performance matches that expected from the original designs. An interesting demonstration of the robustness of the muon systems comes from the reconstruction of muons in heavy-ion collisions. Whereas neither experiment was specifically designed for very high reconstruction efficiency in the very special conditions of heavy-ion collisions, it turns out that they can yield significant physics signals for a few key signatures such as J/ψ and Υ, Υ' production [27].

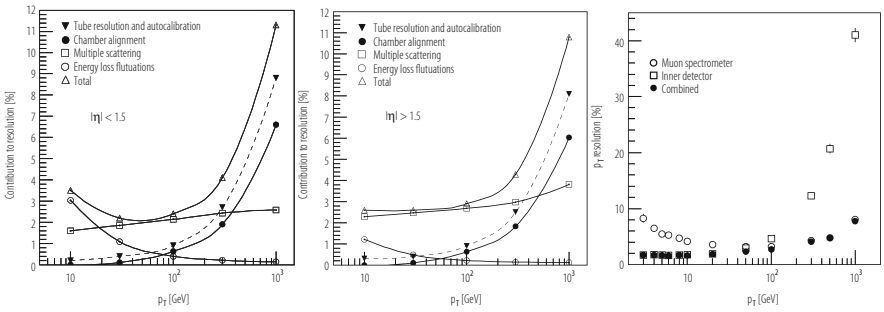


Fig. 16.23 Expected performance of the ATLAS muon measurement. Left: contributions to the momentum resolution in the muon spectrometer, averaged over $|\eta| < 1.5$. Centre: same as left for $1.5 < |\eta| < 2.7$. Right: muon momentum resolution expected from muon spectrometer, inner detector and their combination together as a function of muon transverse momentum

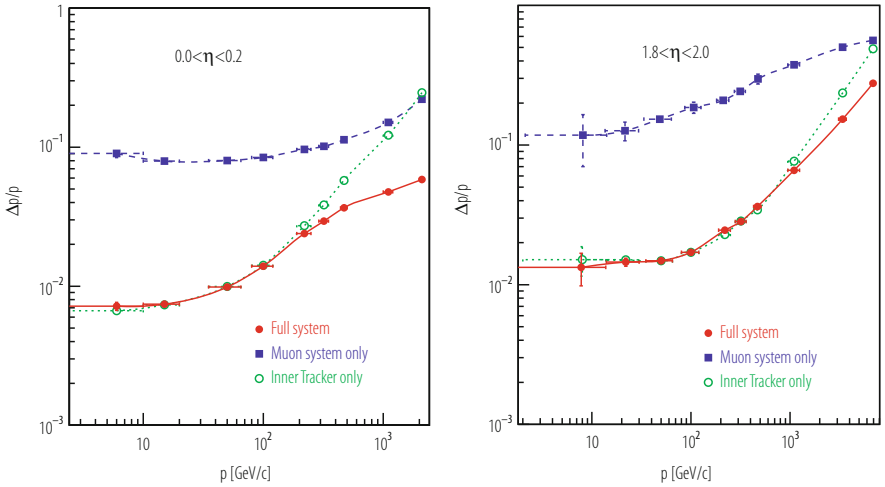


Fig. 16.24 Expected performance of the CMS muon measurement. The muon momentum resolution is plotted versus momentum using the muon system only, the inner tracker only, or their combination (full system) for the barrel, with $|\eta| < 0.2$ (left), and for the end-caps, with $1.8 < |\eta| < 2.0$ (right)

Table 16.13 Main parameters of the ATLAS and CMS muon measurement systems as well as a summary of the expected combined and stand-alone performance at two typical pseudorapidity values (averaged over azimuth)

Parameter	ATLAS	CMS
Pseudorapidity coverage		
Muon measurement	$\|\eta\| < 2.7$	$\|\eta\| < 2.4$
Triggering	$\|\eta\| < 2.4$	$\|\eta\| < 2.1$
Dimensions [m]		
Innermost (outermost) radius	5.0 (10.0)	3.9 (7.0)
Innermost (outermost) disk (z-point)	7.0 (21–23)	6.0–7.0 (9–10)
Segments/super-points per track for barrel (end-caps)	3 (4)	4 (3–4)
Magnetic field B [T]	0.5	2
Bending power (BL [Tm]) at $\|\eta\| \approx 0$	3	16
Bending power (BL [Tm]) at $\|\eta\| \approx 2.5$	8	6
Combined (stand-alone) Momentum resolution at		
$p = 10$ GeV/c and $\eta \approx 0$	1.4% (3.9%)	0.8% (8%)
$p = 10$ GeV/c and $\eta \approx 2$	2.4% (6.4%)	2.0% (11%)
$p = 100$ GeV/c and $\eta \approx 0$	2.6% (3.1%)	1.2% (9%)
$p = 100$ GeV/c and $\eta \approx 2$	2.1% (3.1%)	1.7% (18%)
$p = 1000$ GeV/c and $\eta \approx 0$	10.4% (10.5%)	4.5% (13%)
$p = 1000$ GeV/c and $\eta \approx 2$	4.4% (4.6%)	7.0% (35%)

16.8.4 Trigger Performance

The trigger involves, by design, the selection of only a small fraction of the p−p collisions at the LHC. As a result, a number of compromises on the extent of the physics programme have had to be made. This is an important difference with respect to the experience in e^+e^- machines.

Efficient use of DAQ bandwidth requires that two conditions be fulfilled. First, each level of the trigger attempts to identify physics objects (leptons, photons and jets) as efficiently as possible, while keeping the output bandwidth within requirements. The selected event sample should include all events which would be found by the full offline reconstruction. Hence, the cuts in the trigger must be consistent with those of the offline analysis. Second, since the bandwidth to permanent storage media is limited, events must be selected with care at the final trigger level.

A crucial ingredient of physics analysis is the determination of the trigger efficiency. Three tools allowing the measurement of the requirements imposed by the L1 trigger have been included in the designs. One tool is the presence of overlapping programmable triggers, which allows triggers with different thresholds and cuts to run simultaneously, producing multiple results in parallel. A second tool is prescaled triggers with either lower thresholds or looser requirements (or both) to run in parallel with the main algorithm. A third tool is prescaling of a particular trigger with one of its cuts removed.

Beyond these three tools, another method for measuring the trigger efficiency, which is used extensively, is the use of processes with two physics objects where the trigger selects one of the two. As an example, $Z \rightarrow ee$ decays, selected via the single-electron trigger, can be used to measure the electron trigger efficiency by examining the second, unbiased, electron leg.

A key task is the creation of the trigger tables, i.e. the requirements demanded online, by both the L1 and HLT systems, on the events selected. Table 16.14 lists two examples from ATLAS and CMS, for the L1 trigger. There are, naturally, very significant uncertainties in these rate estimates. At one extreme, CMS allocates only one-third of the assumed DAQ bandwidth to specific triggers. In the ATLAS case, the plan is to absorb any differences in rate via changes in thresholds. Both experiments plan to allocate bandwidth to B physics as well, within the limitations of the total resources available, at the initially low luminosities of the LHC.

The real-time nature of the selections imposes very stringent requirements on the monitoring of the L1 and HLT performance. Initially, many triggers will be run in forced-accept mode, thereby providing the possibility to analyze in detail their performance offline. The trigger monitoring itself will employ a number of tools, including the storage of a small fraction of the events rejected, the comparison of the actual online decisions (as obtained from intermediate hardware calculations that will be stored along with the detector data) and a number of unbiased events, or “minimum-bias” events, which are selected at random, i.e. without any specific requirements on the bunch crossing in question.

Table 16.14 Examples of L1 trigger tables from ATLAS and CMS

	ATLAS		CMS	
Trigger type	Threshold [GeV]	Rate [kHz]	Threshold [GeV]	Rate [kHz]
Inclusive isolated electron/photon	25	12.0	29	3.3
Di-electrons/di-photons	15	4.0	17	1.3
Inclusive isolated muon	20	0.8	14	2.7
Di-muons	6	0.2	3	0.9
Single tau-jet trigger	–	–	86	2.2
Two $\tau-$jets	–	–	59	1.0
Tau-jet * E_T^{miss}	25 * 30	2.0	–	–
1-jet, 3-jets, 4-jets	200, 90, 65	0.6	177, 86, 70	3.0
Jet * E_T^{miss}	60 * 60	0.4	88 * 46	2.3
Electron * Jet	–	–	21 * 45	0.8
Electron * Muon	15 *10	0.1	–	–
Minimum-bias (calibration)			None	0.9
Others (monitor, calibration, ...)		5.0	–	–
Total		25		16

The table corresponds to an instantaneous luminosity of $2 \cdot 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and an assumed total DAQ bandwidth of 25 and 50 kHz respectively. In the case of CMS, only one third of the DAQ bandwidth is allocated, as a safety factor, to account for all the uncertainties in the estimations of the rates. In both cases the threshold corresponds to the point where the efficiency is 95% of the asymptotic efficiency

The trigger systems of the two experiments are also expected to be flexible enough to adapt to changing run and/or coast conditions. As an example, the instantaneous luminosity is expected to drop in the course of a fill, and therefore an optimal allocation of resources might be to change trigger conditions, for instance by lowering trigger thresholds or decreasing pre-scale factors for selected channels. All such changes, along with any other changes in the running conditions, will be logged and the overall online monitoring must record the operational performance as a function of the changes made in real time.

A measure of the performance is given by the efficiency to trigger on single physics objects, namely electrons and photons, muons, jets and tau-jets. The presumed efficiency depends, of course, on the production process and for this reason, Standard-Model processes are used. Table 16.15 lists the efficiencies at L1 and HLT for electrons and muons. For jets, the relevant parameter is not the efficiency which can always reach 100%, but rather the effective threshold needed in order to obtain a fixed efficiency, e.g. 95%, for jets with a certain threshold at the generator level. The situation with τ-jets is more complicated, since the two experiments have studied them in the context of specific physics signatures, which are not directly comparable.

The performance of the L1 trigger and HLT systems has been checked against all the benchmark "major discovery channels" in extensive studies by the two

Table 16.15 Efficiency for triggering on a key physics objects in ATLAS and CMS

Object	ATLAS	CMS
Electrons	$E_T > 25\,\text{GeV}$	$E_T > 29\,\text{GeV}$
L1 efficiency	95%	95%
HLT efficiency	80%	77%
Muons	$P_T > 20\,\text{GeV}$	$P_T > 19\,\text{GeV}$
L1 efficiency	95%	90%
HLT efficiency	80%	77%

The calculations have been performed at different thresholds, which are indicated in the table

experiments. These include all the expected decays of the Standard Model Higgs boson as well as those of the multiple Higgs bosons in the case of supersymmetry. In most cases, the decays involve multiple leptons and can therefore be triggered with very high efficiency. The efficiency to other signatures, such as those expected from supersymmetry is also very high. Overall, current expectations are that the two experiments can address the full physics program that will be made available by the LHC.

16.9 Ten Years of Operation and Physics Analysis in a Nutshell

This section, written 10 years after the previous ones, attempts the impossible, namely to summarise briefly what has been learned at the LHC over the past years. This attempt is limited to the $p-p$ collision data-taking of the ATLAS and CMS experiments, leaving out by necessity entire areas of exciting results obtained in heavy-flavour physics by the LHCb experiment and in heavy-ion physics by ALICE (and also ATLAS and CMS). Most of the examples shown below are taken from ATLAS public results obtained at various stages of the data-taking and physics analysis.

Table 16.16 summarises the different phases of the commissioning and data-taking periods of the ATLAS experiment, as extracted from its already long history of more than 25 years (celebrated in October 2017 in the Bratislava ATLAS week). The first data taken and analysed with the embryonic software under development for the experiment took place in the combined test-beams at the CERN SPS where almost complete slices of the ATLAS detector were exposed to various particle beams over a wide range of energies in the years 2002 to 2006. The next step towards commissioning the experiment took place in the ATLAS cavern itself with combined cosmic runs which illuminated the whole detector, from pixels to outermost muon chambers, and provided a first realistic test-bed for the offline alignment of all sub-systems using the precise measurements of charged-particle tracks in the complex magnetic field of the experiment (silicon sensors, straw tubes, and monitored drift tubes).

Table 16.16 Successive steps in preparation, commissioning, and operation of the ATLAS detector at the LHC

2002 to 2006	Combined test-beams at the CERN SPS
2008 onwards	Combined cosmics
2009	0.9 TeV pp collisions
2010 to 2012	Run-1
2010	7 TeV pp collisions, 36 pb^{-1}
2011	7 TeV pp collisions, 5 fb^{-1}
2012	8 TeV pp collisions, 20 fb^{-1}
2015 to 2018	Run-2
2015	13 TeV pp collisions, 3.2 fb^{-1}
2016	13 TeV pp collisions, 32.8 fb^{-1}
2017	13 TeV pp collisions, 44 fb^{-1}
2018	13 TeV pp collisions, 59 fb^{-1}

The successive years of operation with proton–proton collisions are shown together with the integrated luminosity accumulated each year

16.9.1 Accelerated History: Rediscovering the Standard Model

The first beams at LHC injection energy in 2008 provided huge excitement with only a handful of events called beam splashes produced by single beams interacting in the collimator material just before reaching the experiments. With these events alone, an accurate timing (to $\sim$1 ns) of most of the detector readout channels was achieved, a major step towards commissioning the whole experiment for data-taking with beams. The incident which occurred in the LHC at that point was perceived as a major setback at the time, resulting in a 1 year delay for the LHC to deliver first stable beams with collisions in all experiments. This finally happened in a growing atmosphere of excitement at the end of 2009 at the modest centre-of-mass energy of 0.9 TeV, which corresponds to the injection energy of the proton beams from the CERN SPS into the LHC.

These first few days of data-taking led to the first public results from the LHC experiments and even to a few papers with the first measurements of charged particle multiplicities and differential spectra [34]. The data turned out to be also a wonderful test-bed for rediscovering a large fraction of the very diverse zoo of particles produced in pp interactions. One example is shown in Fig. 16.25 with distinctive peaks at the masses of the π^0 and η mesons in the diphoton spectrum, visible above the combinatorial background from random combinations of pairs of photons reconstructed in the electromagnetic calorimeters.

Another later example of this zoo of particles is shown in Fig. 16.26 based on the first run-2 dataset at 13 TeV from CMS, where one distinguishes clearly among other resonances the narrow J/ψ, Υ, and Z mass peaks used for precise calibration and efficiency measurements of the reconstructed muons across a wide range of energy and pseudorapidities.

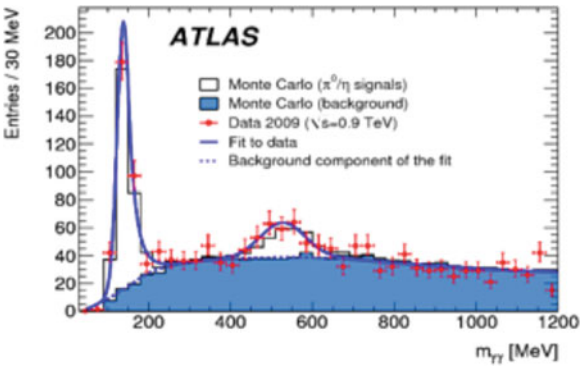


Fig. 16.25 Invariant mass distribution of low-mass diphoton events, as measured in ATLAS with early data at $\sqrt{s} = 0.9\,\text{TeV}$

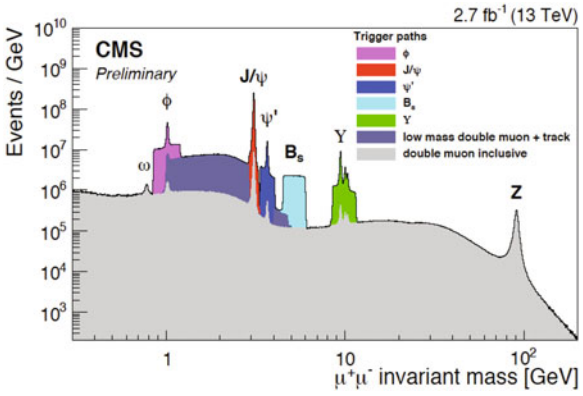


Fig. 16.26 Invariant mass distribution of dimuon events, as measured in CMS with early data at $\sqrt{s} = 13\,\text{TeV}$

In 2010, the very modest accumulated integrated luminosity of $36\,\text{pb}^{-1}$, more than one thousand times smaller than that accumulated in 2017, was nevertheless amply sufficient to observe and measure W/Z-boson production and the production of pairs of top quarks, as shown, respectively, in Figs. 16.27 [35] and 16.28 [36]. Placing LHC measurements on top of the precise predictions from QCD for these production cross-sections as a function of centre-of-mass energy, way beyond previous hadron colliders where these particles were discovered, was the first step in paving the way towards precise tests of the theory with high-statistics measurements based on the very large samples expected in the later years. As of 2019, ATLAS and CMS have accumulated samples of more than 500 million $W \to l\nu$ decays, 50 million $Z \to ll$ decays, and respectively, five million pairs of top quarks with one semi-leptonic top decay and 0.3 million high-purity pairs of top quarks with one electron, one muon, and two b-tagged jets in the final state.

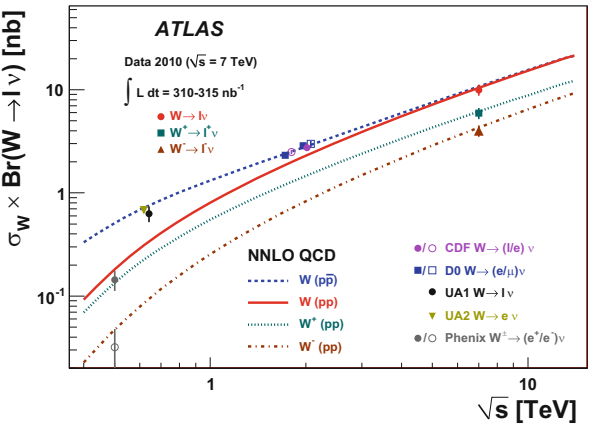


Fig. 16.27 W-boson production cross-section times branching fraction to an electron or muon plus a neutrino, as measured at hadron colliders by PHENIX at RHIC, by UA1/UA2 at the S$p\bar{p}$S, by CDF/D0 at the Tevatron, and by ATLAS at the LHC. The theoretical predictions are shown for both proton-proton and proton-antiproton collisions as a function of the centre-of-mass energy. The ATLAS data correspond to an integrated luminosity of $0.32\,\text{pb}^{-1}$ obtained in 2010 at $\sqrt{s} = 7\,\text{TeV}$

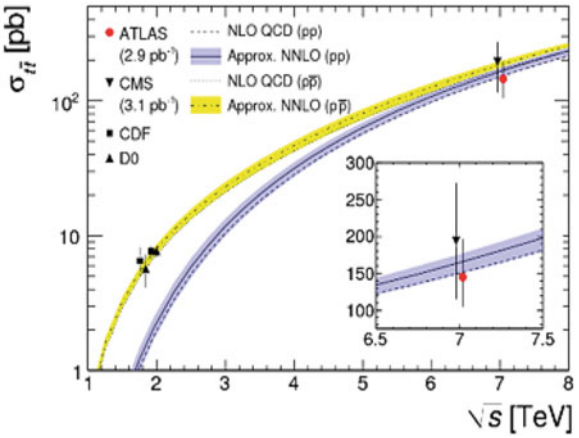


Fig. 16.28 Top quark pair-production cross-section, as measured at hadron colliders by CDF/D0 at the Tevatron and by ATLAS/CMS at the LHC. The theoretical predictions for proton-proton and proton-antiproton collisions assume a top-quark mass of 172.5 GeV and are shown as a function of the centre-of-mass energy. The ATLAS and CMS data correspond to an integrated luminosity of approximately $3\,\text{pb}^{-1}$ obtained in 2010 at $\sqrt{s} = 7\,\text{TeV}$

16.9.2 Precision Measurements

The heavy fundamental particles discussed above are thus an abundant source of prompt isolated electrons and muons, and also, in the case of the Z boson, of hadronically decaying τ-leptons, and have been used extensively in each period

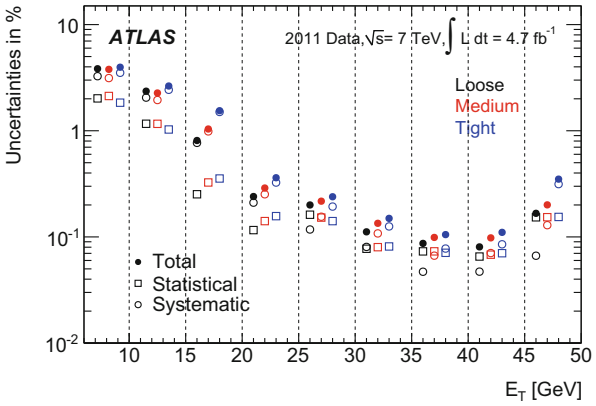


Fig. 16.29 Breakdown of the total uncertainty in the electron combined reconstruction and identification efficiencies, as a function of transverse energy, for the various identification criteria in ATLAS

of data-taking to assess the performance of the detector to reconstruct, identify, and measure their decay products, as well as to provide the most abundant source of triggers for the search for the Higgs boson and for new physics beyond the Standard Model (SM).

Figure 16.29 [37] shows that the efficiencies for reconstructing and identifying prompt isolated electrons could be measured in ATLAS with an overall accuracy ranging from the permil level near the Jacobian peaks from W/Z-boson decays to a few percent in the range 7–10 GeV turned out to be of critical importance for the search for the Higgs boson decaying to four leptons and for still ongoing searches for supersymmetric particles in the electroweak sector.

Figure 16.30 [38] illustrates the calibration accuracy achieved for prompt isolated muons, displayed as a function of the leading muon pseudorapidity for the already very large samples obtained with ATLAS in the run-1 8 TeV data. Tens of millions of J/ψ and Z-boson decays were used to calibrate the data and correct the simulation to reach an overall accuracy at the permil level, leading later on to very precise measurements of the Higgs-boson and W-boson masses. The dimuon events from the intermediate-mass Υ resonance were not used for the calibration itself and served as an independent validation sample to verify the closure of the procedure in terms of its uncertainties.

With sufficiently large samples of prompt isolated electrons, muons and photons, the jets produced in association with these precisely measured objects could be calibrated in situ to a precision far exceeding the initial expectations. Figure 16.31 [39] illustrates this in terms of the overall jet energy scale uncertainty in ATLAS from first run-2 data as a function of jet transverse momentum. The in situ absolute calibration achieves an overall uncertainty at the percent level or even below over a large kinematic range. Uncertainties due to the expected response differences for

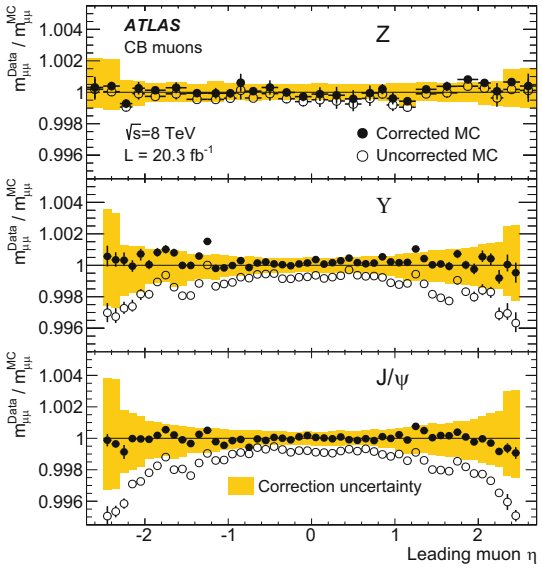


Fig. 16.30 Ratio of the fitted mean mass, $< m_{\mu\mu} >$, for data over simulation (MC), from Z (top), Υ (middle), and J/ψ (bottom) decays to dimuon pairs, as a function of the pseudorapidity of the highest-p_T muon in ATLAS. The ratio is shown for corrected MC (filled symbols) and uncorrected MC (empty symbols). The error bars represent the overall statistical and systematic uncertainty obtained from the mass fits. The bands show the uncertainties in the MC corrections calculated separately for the three samples

quark versus gluon jets and to pile-up at low transverse momenta dominate however the overall uncertainty on the jet energy scale over most of the range.

Precisely measured objects in simple final states lead to precisely measured fiducial differential and integrated cross-sections, which can then be compared to state-of-the-art theoretical predictions and used for example to improve the uncertainties in the parton distribution functions in the proton. Two examples of such ATLAS measurements, among the most precise to-date at the LHC, are shown as an illustration in Figs. 16.32 [40] and 16.33 [41], for inclusive jets as a function of jet transverse momentum in different rapidity ranges and for the integrated $W^{\pm}$ versus Z/γ^* cross-sections, respectively.

These precision measurements together with a wealth of others are not only used to improve the knowledge of the parton distribution in the proton, but also to improve the theoretical modelling of the relevant production processes, thereby reducing theoretical uncertainties which today are dominant when considering the measurement of fundamental Standard Model parameters such as the W-boson mass and the weak mixing angle.

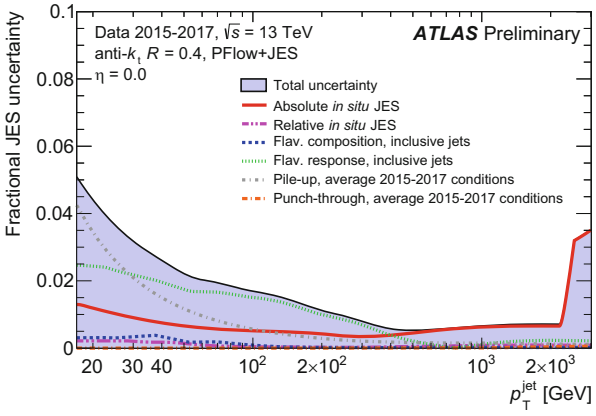


Fig. 16.31 Fractional jet energy scale (JES) systematic uncertainty components as a function of jet transverse momentum, p_T for jets reconstructed at central pseudorapidity from particle flow objects in ATLAS. The total uncertainty (all components summed in quadrature) is shown as a filled region topped by a solid black line. Topology-dependent components are shown under the assumption of a dijet flavour composition. At values of p_T, the uncertainty from the pile-up of $p-p$ interactions in the same or neighbouring bunch-crossings dominates the overall jet energy scale uncertainty. The data shown represent an average over the run-2 period from 2015 to 2017, corresponding to an average number of 30 interactions per bunch crossing

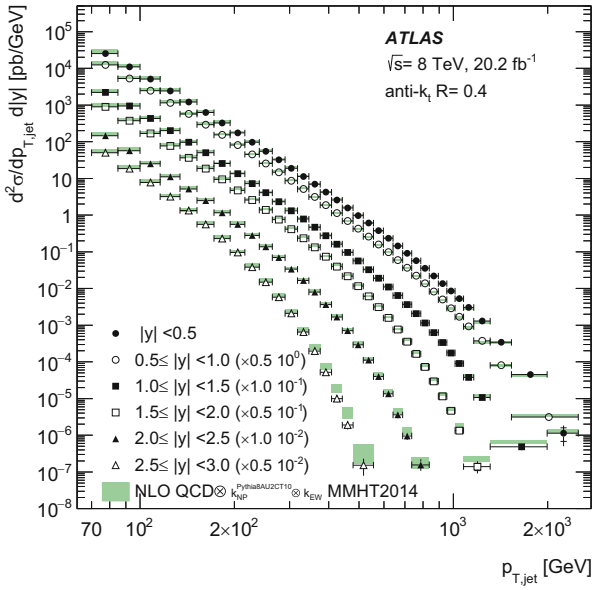


Fig. 16.32 Inclusive jet cross-section as a function of jet transverse momentum, p_T, in bins of jet rapidity. The results are shown for standard jets as measured with ATLAS 8 TeV data. The data are compared to the next-to-leading order QCD predictions with the MMHT2014 parton distribution function set, corrected for non-perturbative and electroweak effects

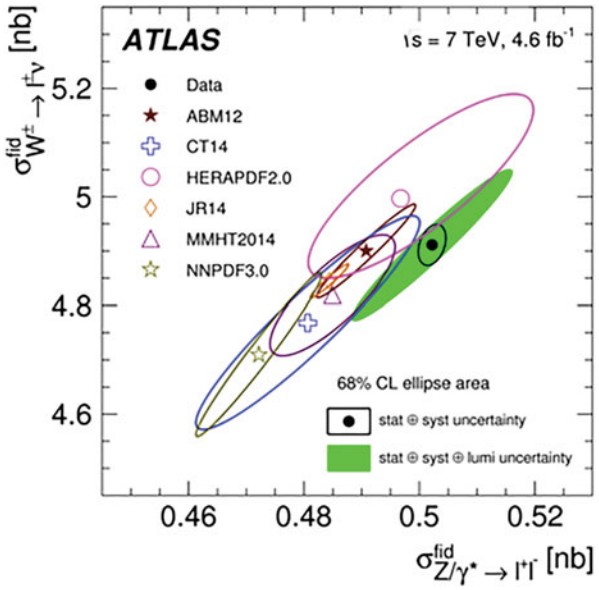


Fig. 16.33 Integrated fiducial cross sections times leptonic branching fractions, σ_W^{fid} versus σ_Z^{fid}, as measured with ATLAS 7 TeV data. The data ellipses display the 68% confidence level coverage for the total uncertainties (full green) and total excluding the luminosity uncertainty (open black). Theoretical predictions based on various parton distribution function (PDF) sets are shown with open symbols of different colours. The uncertainties of the theoretical calculations correspond to the PDF uncertainties only

16.9.3 Discovery and Measurements of the Higgs Boson

The search for the Higgs boson, over a wide mass range, was a major goal and challenge for the LHC physics programme, and the expected signatures from Higgs-boson decays therefore served as benchmarks to optimise the detector design from the very beginning in the late 1980's. These signatures span the full range of physics objects which can be reconstructed, identified and measured precisely in the experiments. The four-lepton $H \to ZZ^* \to 4l$ and the dillepton plus missing transverse energy $H \to WW^* \to l\nu l\nu$ channels were expected to be the most sensitive ones for Higgs-boson masses above 120–130 GeV. For lower values of the Higgs-boson mass, as favoured by the combined precision electroweak fits to the data available before LHC turn-on, the diphoton channel $H \to \gamma\gamma$ channel was expected to be the most sensitive channel.

The expectations for Higgs-boson discovery in the 1990's required integrated luminosities of approximately 30 to 100 fb^{-1} at the nominal LHC centre-of-mass energy of 14 TeV for Higgs-boson discovery in a single decay channel. These were updated before LHC operation with more precise theoretical calculations, resulting in particular in a significant increase of the dominant Higgs-boson

production cross-section through gluon-gluon fusion, to simple combinations of the most sensitive channels, and finally to the reduced 7 TeV centre-of-mass energy of the initial run-1 data. These updated expectations, leading to potential discovery with as little as 5–10 fb^{-1} of integrated luminosity, resulted in a period of great excitement within the ATLAS and CMS experiments, but also in the community at large, from summer 2011 (with 1 fb^{-1} collected by the experiments) to summer 2012 when the Higgs boson was officially announced as having been discovered by each of the two experiments. The evolution of the Higgs-boson signal significance over this period is illustrated in Fig. 16.34. In summer 2011, as shown in Fig. 16.34a, there were no indications of any signal yet and the fluctuations observed as a function of mass were compatible with background fluctuations. At the end of 2011, however, both experiments had excluded a Standard Model Higgs-boson signal over a mass range extending from the LEP limit of 114 to 600 GeV, except for a narrow mass range around 125 GeV in which the largest deviation from background expectations was observed around 125 GeV and corresponded to approximately three standard deviations in each experiment, as shown in Fig. 16.34b. Finally, Fig. 16.34c,d shows the observed significance in summer 2012 when the discovery was claimed and subsequently published by both experiments [42, 43] for 10 fb^{-1} of data at 7 and 8 TeV.

The four-lepton and diphoton channels have always been rightly considered as the two best channels for Higgs-boson discovery, since they both provide a clear and narrow peak for the Higgs-boson signal in the invariant mass distribution of the final state particles on top of a continuous background. In addition the four-lepton channel can be observed above a much smaller continuum background, consisting predominantly of continuum $ZZ* \rightarrow 4l$ final states. These features can be seen in Figs. 16.35 and 16.36 taken from the ATLAS discovery publication [42]. In contrast, the third channel which contributed to the discovery, namely the $H \rightarrow WW* \rightarrow l\nu l\nu$ channel, has a poor mass resolution because of the presence of neutrinos in the final state, as shown in Fig. 16.37.

After the discovery, measurements of the properties of the Higgs boson were performed in successive stages, first focusing on its spin, then on its couplings to bosons and fermions and on possible non-SM contributions to its width. At the end of run-1, ATLAS and CMS produced a combined paper on the Higgs-boson couplings [44], leading to the conclusion that in all production modes and decay channels which had been measured at the time, the Higgs-boson properties were compatible with what one would expect from the SM. More recently, each experiment has produced updated results based also on a large fraction of the run-2 data. This is illustrated in Fig. 16.38, which is based on the most recent run-2 ATLAS Higgs combination results [45] and shows that the strength of the measured Higgs-boson couplings to fermions and bosons follows the expectations from the SM, in which for example the Yukawa fermion coupling is expected to be proportional to the fermion mass. Finally, based on the most recent results from the combined run-1 and run-2 datasets from ATLAS and CMS [46], Table 16.17 shows

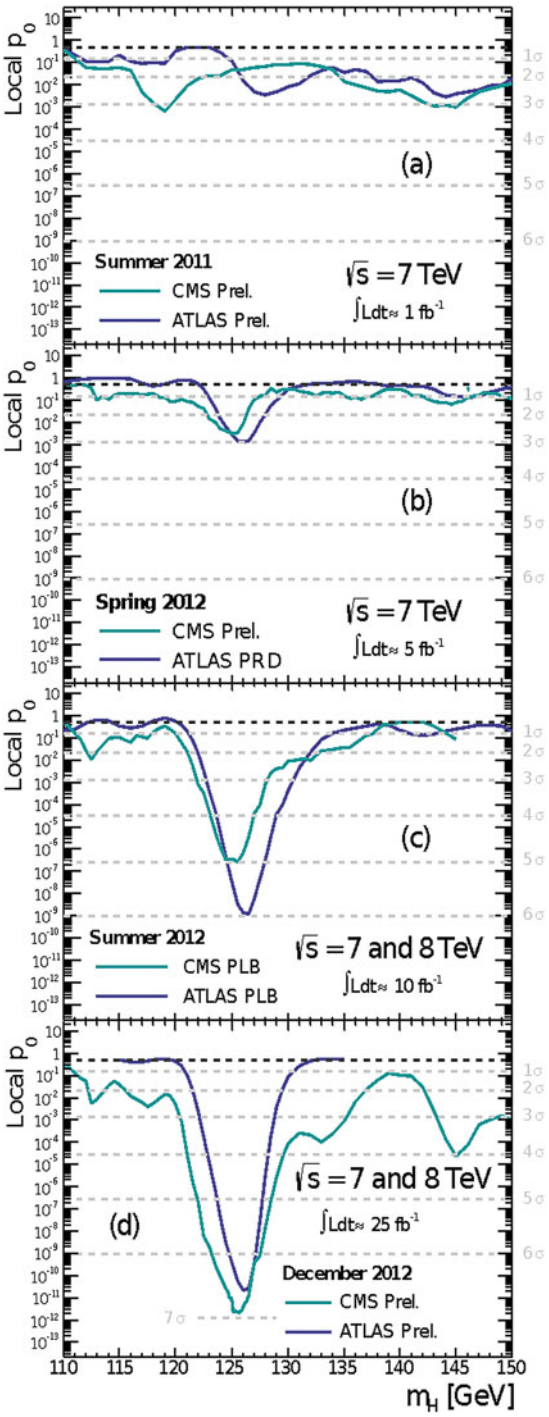


Fig. 16.34 Evolution of the combined significance of the Higgs-boson signal in the ATLAS and CMS experiments from exclusion limits in summer 2011 to discovery in summer 2012

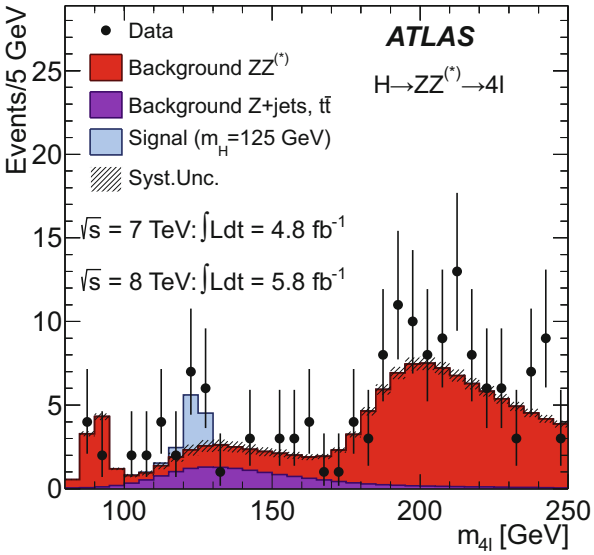


Fig. 16.35 Distribution of the four-lepton invariant mass for the selected candidates in the $H \rightarrow ZZ^* \rightarrow 4l$ channel, as observed by ATLAS at the time of discovery in summer 2012. The expected signal for $m_H = 125\,\text{GeV}$ is shown stacked on top of the overall background prediction

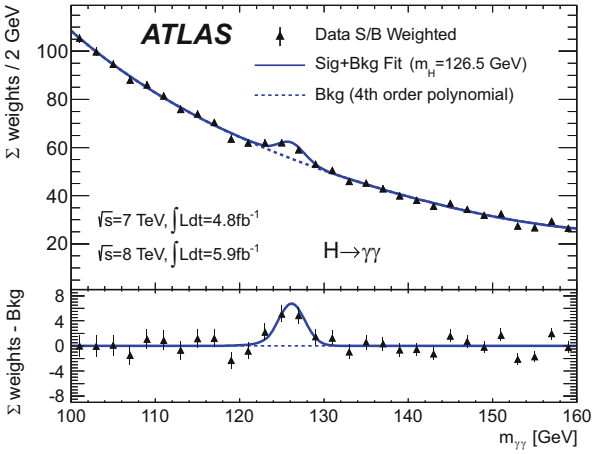


Fig. 16.36 Distribution of the invariant mass of diphoton candidates in the $Htoγγ$ channel, as observed by ATLAS at the time of discovery in summer 2012. The expected signal for $m_H = 125\,\text{GeV}$ is shown stacked on top of the overall background prediction. The residuals of the weighted data with respect to the fitted background is displayed in the bottom panel

that the Higgs couplings to charged third-generation fermions are now all clearly observed unambiguously and measured to be compatible with SM expectations. In contrast to the channels used for the discovery, the vast majority of the signals

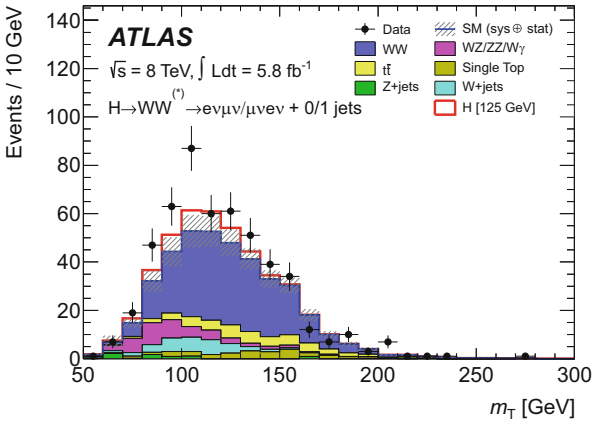


Fig. 16.37 Distribution of the transverse mass of the Higgs boson candidates in the $H \to WW$ decay channel, as observed by ATLAS at the time of discovery in summer 2012. The expected signal for $m_H = 125\,\text{GeV}$ is shown stacked on top of the overall background prediction

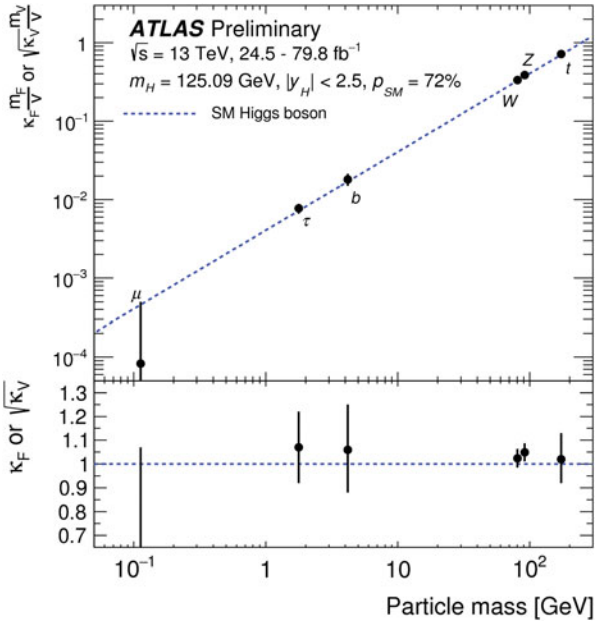


Fig. 16.38 Reduced coupling strength modifiers $\kappa_F m_F/v$ for fermions ($F = t, b, \tau, \mu$) and $\sqrt{\kappa_V} m_V/v$ for weak gauge bosons ($V = W, Z$) as a function of their masses m_F and m_V, respectively, where the vacuum expectation value of the Higgs field $v = 246\,\text{GeV}$. The results are obtained from ATLAS 13 TeV data and the SM prediction is also shown (dotted line). The coupling modifiers κ_F and κ_V are measured assuming that there are no beyond-SM contributions to the Higgs-boson decays or production processes. The lower inset shows the ratios of the measured values to their SM predictions

Table 16.17 Summary of direct measurement of all Yukawa couplings of the Higgs boson to third-generation charged fermions (τ lepton, bottom quark, and top quark) shown for the ATLAS and CMS experiments

		τ lepton	Bottom quark	Top quark
ATLAS	Observed significance	6.4σ	5.4σ	6.3σ
	Expected significance	5.4σ	5.5σ	5.1σ
	Measured to predicted yield ratio	1.09 ± 0.35	1.01 ± 0.20	1.34 ± 0.21
CMS	Observed significance	5.9σ	5.5σ	5.2σ
	Expected significance	5.9σ	5.6σ	4.2σ
	Measured to predicted yield ratio	1.09 ± 0.29	1.04 ± 0.20	1.26 ± 0.28

The expected and observed signal significances are listed, together with the ratios of the observed yields to those predicted by the SM

explored in these cases are among the most difficult Higgs-boson measurements due to the diverse and potentially large backgrounds and to the fact that the signal does not yield a narrow peak above the background.

16.9.4 Search for New Physics: Dashed and Renewed Hopes

The search for signatures from new physics beyond the SM has been ongoing in many directions from the very beginning of LHC data-taking, as has always been the case when an accelerator at the energy frontier begins operation and almost immediately delivers data to the experiments which allow them to supersede the limits from previous searches very quickly in certain cases, such as those obtained at the Tevatron. In the early years of data-taking, the experimental analyses were very much geared towards discovery because each year of data-taking brought either a large increase in integrated luminosity or a significant boost in centre-of-mass energy which is the key to searches at the edge of the available phase space. Examples of such searches are shown in Figs. 16.39 and 16.40, based on very recent results from ATLAS.

Figure 16.39 presents the evolution of the limits set by successive ATLAS searches for one of the simplest signatures of new physics, namely that for a new neutral vector boson, Z', decaying into electron or muon pairs. The limit of $\sim$1 TeV on the mass of the Z' boson in the case of a simple sequential extension of the SM was already competitive in 2010 with the legacy search limits from the CDF/D0 experiments at the Tevatron. With the full run-2 dataset, the limit is now set at 5 TeV [47] and will not extend much further without any further increase of the beam energy. Figure 16.40 shows a similar evolution of the limits set on possible excited quarks decaying into a pair of high transverse momentum jets [48].

Since 2017, however, these golden years for the excitement of searches at the edge of the available phase space are gone, and the focus of the analyses has been more on the more difficult and exotic signatures of new physics. In particular, despite its theoretical beauty before symmetry breaking, supersymmetry, if realised

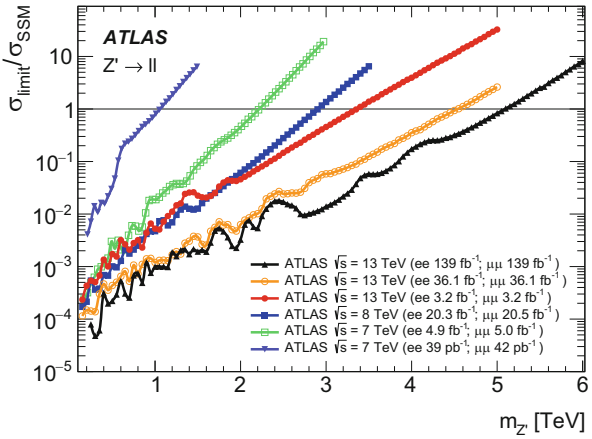


Fig. 16.39 Ratio of the observed cross-section limit to the expected Z' cross-section in the Sequential Standard Model for the combination of the dielectron and dimuon channels. The ratio is shown as a function of the Z' mass for a number of ATLAS searches performed at various LHC centre-of-mass energies from 2010 to 2018

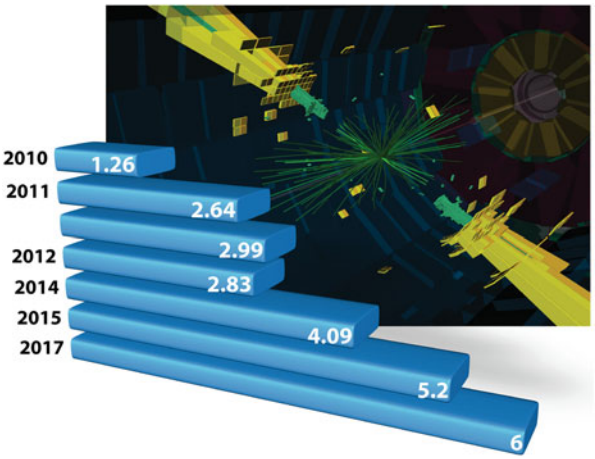


Fig. 16.40 Evolution of exclusion limits in TeV set by ATLAS on dijet resonance searches, interpreted as arising from the decay of an excited quark, from 2010 to 2017. The background image shows a display of one of the highest-mass ATLAS dijet events

in nature, has remained elusive and beyond the reach of the experimental searches in even the most exotic scenarios envisaged for its possible manifestation at the scales at which it is probed. In most models, the third generation supersymmetric partners of the quarks, the so-called stop quarks, are expected to have the smallest mass and therefore to be the most accessible at the LHC. Since their decay signatures involve predominantly top and bottom quarks, the search for these particles has had

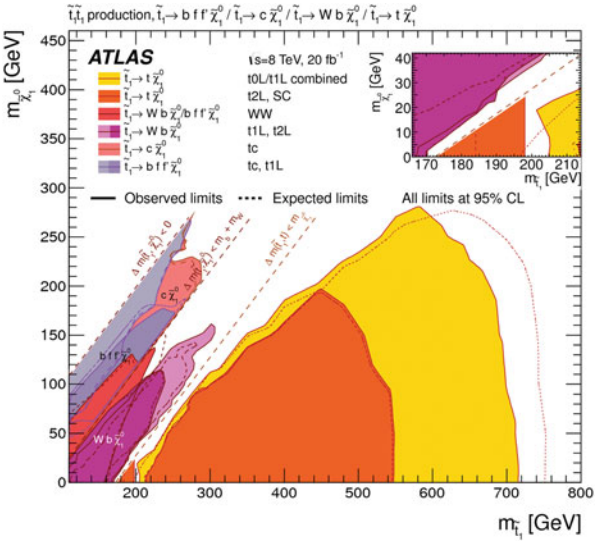


Fig. 16.41 First summary plot based on ATLAS run-1 data at $\sqrt{s} = 7$ and 8 TeV on searches for top squarks, showing the top squark versus lightest supersymmetric particle mass plane

to branch into many complex signatures, leading at first to only a partial coverage of the accessible parameter space in terms of the masses of the lightest stop quark and of the lightest neutralino, assumed to be stable. This is illustrated in Fig. 16.41, based on ATLAS run-1 data [49]. The sensitivity at the time reached at best a mass of 700 GeV and the searches were not yet very sensitive to stop quark masses close to the top-quark mass itself. Eight years later, after several generations of ever more complex and diverse searches for the stop quark, Fig. 16.42 shows that the sensitivity has extended to masses close to 1000 GeV [49], and that most of the plane of possible masses is now excluded for a lightest neutralino mass below 300 GeV.

Perhaps the most striking example of the huge efforts put by ATLAS and CMS into hunting supersymmetry has been the search for the weakly interacting supersymmetric particles, with names such as chargino, neutralino, slepton or Higgsino. It has taken the LHC experiments much longer to supersede the limits from the experiments at the LEP electron-positron collider for some of these hypothetical supersymmetric particles because of the small cross-sections involved and of the rather low energies of the decay products, leading therefore to potentially large backgrounds from SM processes with similar signatures and much larger cross-sections. This is illustrated in Fig. 16.43 which presents the most recent limits on the heavier chargino and neutralino masses as a function of the lightest neutralino mass for cases where the lightest neutralino is assumed to be stable [49].

The few results shown here, together with, for example, the very active ongoing searches for dark matter or long-lived particles, demonstrate that there are many areas still to be covered in the search for new physics at the LHC. The accelerator

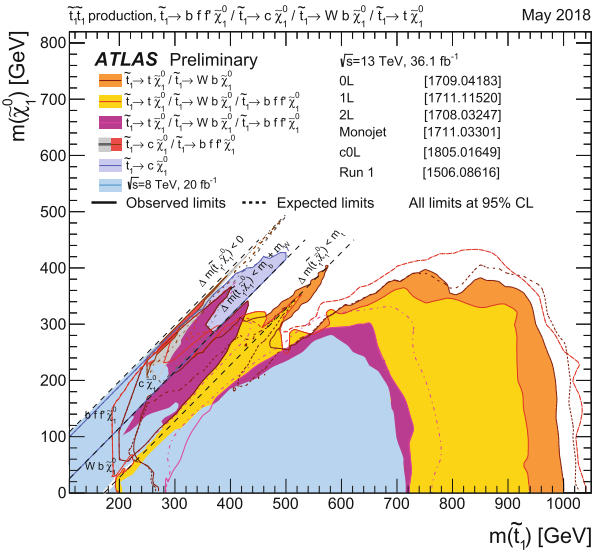


Fig. 16.42 Summary plot based on ATLAS 2015-2016 data at $\sqrt{s} = 13\,\text{TeV}$ on searches for top squarks, showing the top squark versus lightest supersymmetric particle mass plane

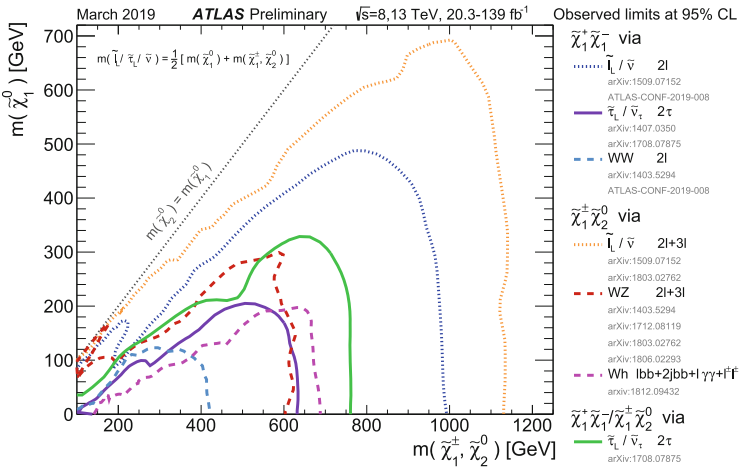


Fig. 16.43 For a variety of ATLAS datasets and search channels, 95% confidence-level exclusion limits on supersymmetric neutralino and chargino production as a function of their mass versus that of the lightest supersymmetric particle (assumed to be stable). Each individual exclusion contour represents one or more analysis in simple merged curves

and all its experiments will remain for many many years to come a wonderful provider of new data in this quest for physics beyond the Standard Model, however elusive it may be.

16.10 Conclusion

The formidable challenge related to the design, construction, installation, and commissioning of the ATLAS and CMS experiments reached a successful conclusion at the end of 2009 with the beginning of data-taking. At the time, the next challenge was as daunting and even more exciting for all the physicists participating in the exploitation phase: understand the performance of these unprecedented detectors as precisely as possible and extract the rich harvest of physics, which would undoubtedly show up once the LHC machine achieved its design goals at high energy and high luminosity.

Ten years later, after taking large amounts of data at centre-of-mass energies of 7, 8 and 13 TeV and operating successfully at luminosities exceeding even the design goals of the machine and the experiments, one can look back with tremendous pride and respect at what has been achieved by the thousands of people involved in the accelerator and the experiments. But we have also been very lucky and should feel huge gratitude towards nature which has offered the ATLAS and CMS experiments the possibility to first observe and later measure the Higgs boson in the somehow miraculous variety of production processes and decay channels with which it manifests itself at the LHC. The searches for new physics at this new frontier have, however, unfortunately not yielded yet any sign of where the solutions of some of the remaining mysteries of nature might lie. Nevertheless, the physics harvest already available from this wonderful tool for fundamental research is already rich beyond belief and the ongoing analyses in the experiments continue to probe the Standard Model predictions to the utmost of our current capabilities. Might new physics still emerge from the expected thirty times larger datasets to be collected over the coming ten to 15 years from the upgraded machine and experiments? The hopes remain high, yet only nature knows.

Acknowledgements The author wishes to thank deeply P. Sphicas, with whom the review article of ref. [4] was written. Much of the contents of this chapter are drawn from that review. May all the ATLAS and CMS colleagues who helped in a significant way to prepare this review find here also the expression of the author's most sincere thanks. It would have been impossible to collect the information in this chapter without the help of many experts in a variety of fields across all the aspects of the design, construction and installation of these very large state-of-the-art experiments in particle physics.

References

1. ATLAS Collaboration, *ATLAS Technical Proposal*. CERN/LHCC/94-43 (1994); CMS Collaboration, *CMS Technical Proposal*. CERN/LHCC/94-38 (1994).
2. ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*. JINST 3 S08003 (2008).
3. CMS Collaboration, *The CMS Experiment at the CERN LHC*. JINST 3 S08004 (2008).
4. Froidevaux, D., Sphicas, P., Annu. Rev. Nucl. Part. Sci. **56** (2006) 375–440.

5. ALICE Collaboration, *The ALICE Experiment at the CERN LHC*. JINST 3 S08002 (2008).
6. LHCb Collaboration, *The LHCb Detector at the LHC*. JINST 3 S08005 (2008).
7. Grunewald, M.W., Invited Talk at EPS Int. Europhysics Conf. on High Energy Physics (HEP-EPS-2005), Lisbon, Portugal (2005). arXiv:hep-ex/0511018.
8. Chanowitz, M.S., Phys. Rev. D **66** (2002) 073002. arXiv:hep-ph/0207123.
9. Barbieri, R., Strumia, A., Phys. Lett. B **462** (1999) 144. arXiv:hep-ph/9905281.
10. Birkedal, A., Matchev, K., Perelstein M., arXiv:hep-ph/0412278.
11. Gianotti, F., Mangano, M., CERN Preprint, CERN-PH-TH/2005-072 (2005). arXiv:hep-ph/0504221.
12. Blaising, B., et al., Contribution to the Zeuthen Briefing Book. http://council-strategygroup.web.cern.ch/council-strategygroup/
13. ATLAS Collaboration, *Detector and Physics Performance Technical Design Report*. CERN/LHCC/99-15 (1999);
Asai, S., et al., Eur. Phys. J. C **32** (2004) 19.
14. ATLAS Collaboration, *Magnet System Technical Design Report*. CERN/LHCC/97-18 (1997);
ATLAS Collaboration, *Barrel Toroid Technical Design Report*. CERN/LHCC/97-19 (1997);
ATLAS Collaboration, *End-cap Toroid Technical Design Report*. CERN/LHCC/97-20 (1997);
ATLAS Collaboration, *Central Solenoid Technical Design Report*. CERN/LHCC/97-21 (1997);
CMS Collaboration, *The Magnet Project - Technical Design Report*. CERN/LHCC/97-10 (1997).
15. ATLAS Collaboration, *Technical Co-ordination Technical Design Report*. CERN/LHCC/99-01 (1999).
16. Akesson, T., et al., *Report of the High-Luminosity Study Group to the CERN Long-Range Planning Committee*, ed. Mulvey, J., CERN Yellow Book, 88-02 (1988).
17. ATLAS Collaboration, *Inner Detector Technical Design Report*, Vol. II. CERN/LHCC/97-16. ISBN 92-9083-103-0 (1997);
ATLAS Collaboration, *Pixel DetectorTechnical Design Report*. CERN/LHCC/98-13, (1998).
18. Lindstrom, G., *Radiation damage in silicon detectors*, Nucl. Instrum. Meth. A **512** (2003) 30–43.
19. Coe, P.A., Howell, D.F., Nickerson, R.B., *Frequency scanning interferometry in ATLAS: remote, multiple, simultaneous and precise distance measurements in a hostile environment*. Meas. Sci. Technol. (2004) 2175–2187.
20. Buttar, C.M., et al., Nucl. Instrum. Meth. A **447** (2000) 126;
Dierlamm, A., Nucl. Instrum. Meth. A **514** (2003) 167.
21. Akesson, T., et al., Nucl. Instrum. Meth. A **522** (2004) 25;
Capeans, M., et al., IEEE Trans. Nucl. Sci. **51** (2004) 960;
Akesson, T., et al., Nucl. Instrum. Meth. A **515** (2003) 166;
Romaniouk A., ATLAS Internal Note, ATL-INDET-98-211 (1998).
22. Gorelov, I., et al., Nucl. Instrum. Meth. A **481** (2002)204;
Alimonti, G., et al., ATLAS Internal Note, ATL-INDET-INT-2005-006 (2005);
Alimonti, G., et al., ATLAS Internal Note, ATL-INDET-INT-2005-007 (2005).
23. ALEPH Collaboration, Nucl. Instrum. Meth. A **360** (1995) 481;
OPAL Collaboration, OPAL Technical Note, OPAL-TN-306 (1995).
24. Drage, L., Parker, M.A., ATLAS Internal Note, ATL-PHYS-2000-007 (2000).
25. ATLAS Collaboration, *Liquid Argon Calorimeter Technical Design Report*. CERN/LHCC/96-41 (1996).
26. Colas, J., et al., Nucl. Instrum. Meth. A **550** (2005) 96.
27. CMS Collaboration, *CMS Physics Technical Design Report*, Vol. I: *Detector Performance and Software*. CERN/LHCC/2006-01 (2006).

28. Braunschweig, H., et al., Nucl. Instrum. Meth. A **265** (1988) 246;
Andrieu, B., et al., DESY Preprint, DESY 93-04 (1993).
29. ATLAS Collaboration, *Muon Spectrometer Technical Design Report*.
CERN/LHCC/97-22 (1997).
30. ATLAS Collaboration, *Computing Technical Design Report*. CERN/LHCC/2005-022 (2005).
31. Chytracek, R., et al., Nucl. Science Symp. Conf. Record, IEEE **4** (2004) 2077;
see also http://lcgapp.cern.ch/project/persist/
32. Brun, R., Rademakers, F., Nucl. Instrum. Meth. A **389** (1997) 81;
see also http://root.cern.ch/
33. The LHC Computing Grid. *Technical Design Report*. CERN/LHCC/2005-024 (2005).
34. ALICE Collaboration, Eur. Phys. J **C65** (2010) 111;
ALICE Collaboration, Phys. Lett. **B693** (2010) 53;
CMS Collaboration, JHEP **02** (2010) 041;
ATLAS Collaboration, Phys. Lett. **B688** (2010) 21.
35. ATLAS Collaboration, JHEP **12** (2010) 060.
36. ATLAS Collaboration, Eur. Phys. J. **C71** (2011) 1577.
37. ATLAS Collaboration, Eur. Phys. J. **C74** (2014) 2941.
38. ATLAS Collaboration, Eur. Phys. J. **C74** (2014) 3130.
39. ATLAS Collaboration, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/JETM-2018-006.
40. ATLAS Collaboration, JHEP **09** (2017) 020.
41. ATLAS Collaboration, Eur. Phys. J. **C77** (2017) 367.
42. ATLAS Collaboration, Science **338** (2012) 1576.
43. CMS Collaboration, Science **338** (2012) 1569.
44. ATLAS and CMS Collaborations, JHEP **08** (2016) 045.
45. ATLAS Collaboration, ATLAS-CONF-2019-005 (2019), https://cds.cern.ch/record/2668375.
46. M. Kado, talk given at the Aspen 2019 Winter Conference in Physics, https://indico.cern.ch/event/748043/contributions/3313769/attachments/1817533/2971936/AspenHiggs.pdf.
47. ATLAS Collaboration, CERN-EP-2019-030 (2019), http://arxiv.org/abs/arXiv:1903.06248, submitted to Phys. Lett. B.
48. ATLAS Collaboration, ATLAS-CONF-2019-007 (2019), http://cdsweb.cern.ch/record/2668385.
49. ATLAS Collaboration, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CombinedSummaryPlots/SUSY.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 17
Neutrino Detectors Under Water and Ice

Christian Spiering

17.1 Introduction

Underwater/ice neutrino telescopes are multi-purpose detectors covering astrophysical, particle physics and environmental aspects [1–3]. Among them, the detection of the feeble fluxes of astrophysical neutrinos which should accompany the production of high energy cosmic rays is the clear primary goal [4, 5]. Since these neutrinos can escape much denser celestial bodies than light, they can trace processes hidden to traditional astronomy. Different to gamma rays, neutrinos provide incontrovertible evidence for hadronic acceleration. On the other hand, their extremely low interaction cross section makes their detection extraordinarily difficult.

Figure 17.1 shows a compilation of the spectra of dominant natural and artificial neutrino fluxes. Solar neutrinos, burst neutrinos from SN-1987A, reactor neutrinos, terrestrial neutrinos from radioactive decay processes in the Earth and neutrinos generated in cosmic ray interactions in the Earth atmosphere ("atmospheric neutrinos") have been already detected. Two guaranteed—although not yet detected—fluxes are the diffuse flux of neutrinos from past supernovae (marked "background from old supernovae") and the flux of neutrinos generated in collisions of ultra-energetic protons with the 3 K cosmic microwave background [6] (marked GZK after Greisen, Zatsepin and Kuzmin [7] who first considered such collisions). These neutrinos will hopefully be detected in the next decade. Neutrinos in the TeV-PeV range emerging from acceleration sites of cosmic rays (marked AGN after "Active Galactic Nuclei") have been detected in 2013 with IceCube [8]. No practicable idea exists how to detect 1.9 K cosmological neutrinos.

C. Spiering (✉)
DESY, Zeuthen, Germany
e-mail: christian.spiering@desy.de

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_17

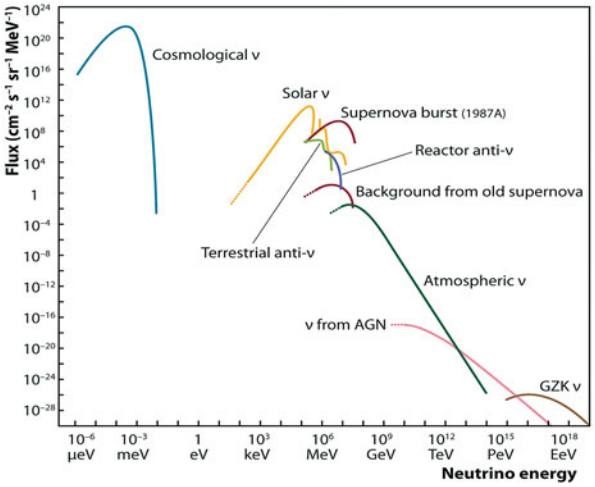


Fig. 17.1 Spectra of natural and reactor neutrinos

The energy range below 5 GeV is the clear domain of underground detectors, notably water Cherenkov, liquid scintillator and radio-chemical detectors (see chapter C4 and [9]) which led to the discovery of solar and atmospheric neutrinos, of neutrino oscillations and of neutrinos from Supernova SN1987A. These detectors, presently with maximal geometrical cross sections of about 1000 m^2, have turned out to be too small to detect the feeble fluxes of astrophysical neutrinos from cosmic acceleration sites. The high energy frontier of TeV and PeV energies is being tackled by much larger, expandable detectors installed in open water or ice, a principle first proposed by M. Markov in 1959 [10]. They consist of arrays of photomultipliers recording the Cherenkov light from charged particles produced in neutrino interactions. Towards even higher energies, novel detectors aim at detecting the coherent Cherenkov radio signals (ice, salt) or acoustic signals (water, ice, salt) from neutrino-induced particle showers. Air shower detectors search for showers with a "neutrino signature". The very highest energies are covered by balloon-borne detectors recording radio emission in terrestrial ice masses, by ground-based radio antennas sensitive to radio emission in the moon crust, or by satellite detectors searching for fluorescence light or radio signals from neutrino-induced air showers. This article focuses on optical detectors in water and ice. The methods for higher energies are sketched in Sect. 17.8. Table 17.1 gives an overview over past, present and future optical detectors in water and ice.

Underwater/ice detectors—apart from searching for neutrinos from cosmic ray sources—also address a variety of particle physics questions (see for reviews [12, 13]). With their huge event statistics, large neutrino telescopes have opened a new perspective for oscillation physics with atmospheric neutrinos, and actually can compete with accelerator experiments [14]. Another example for a particle physics task is the search for muons produced by neutrinos from dark matter annihilation

Table 17.1 Past, present (2018) and future neutrino telescope projects and their main parameters

Experiment	Location	Size (km^3)	Milestones	Remarks
DUMAND	Hawaii		1978/–/–/1996	Terminated due to techn./funding problems
NT200	Lake Baikal	10^{-4}	1980/1993/1998/2015	First proof of principle
NESTOR	Med. Sea off Peloponnes		1991/–/–/–	Data taking with prototype
NEMO	Med. Sea off Sicily		1998/–/–/–	R&D project prototype tests
AMANDA	South Pole	0.015	1990/1996/2000/2009	First deep-ice ν telescope
ANTARES	Med. Sea off Toulon	0.010	1997/2006/2008/2018	First deep-sea ν telescope
IceCube	South Pole	1.0	2011/2005/2010/–	First km^3-sized detector
GVD-1	Lake Baikal	0.4	2012/2015/–/–	High-energy ν astronomy
KM3NeT/ARCA	Med. Sea off Sicily	1–1.5	2013/2015/–/–	High-energy ν astronomy
KM3NeT/ORCA	Med. Sea off Toulon	0.003	2014/2017/–/–	Low-energy configuration for oscillation physics
GVD-2	Lake Baikal	1.5	2012/–/–/–	Extension of GVD-1
KM3NeT Phase 3	Med. Sea	3–5	2013/–/–/–	Planned extension of of KM3NeT
IceCube-Gen2	South Pole	5–10	2014/–/–/–	Planned IceCube extension covering low/high energies, a surface array and radio detection

The milestone years give times of project start, of first data taking with partial configurations, of detector completion, and of project termination. Projects with first data expected past 2025 are in italics (modified after [11])

in the Sun or in the center of the Earth. These searches are sensitive to super-symmetric WIMPs (Weak Interacting Massive Particles) as dark matter candidates. Underwater/ice detectors can also search for relativistic magnetic monopoles, with a light emission 8300 times stronger than that of a bare muon and therefore providing a very clear signature. Other tasks include the search for super-heavy particles like GUT monopoles, super-symmetric Q-balls or nuclearites which would propagate with less than a thousandth of the speed of light and emit light by heating up the medium or by catalyzing proton decays.

The classical operation of neutrino telescopes underground, underwater and in deep ice is recording upward travelling muons generated in a charged current neutrino interaction. The upward signature guarantees the neutrino origin of the muon since no other particle can cross the Earth. A neutrino telescope should be

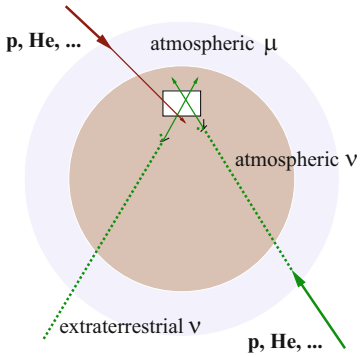


Fig. 17.2 Sources of muons in deep underwater/ice detectors. Cosmic nuclei—protons(p), α-particles(He), etc.—interact in the Earth atmosphere (light-colored). Sufficiently energetic muons produced in these interactions ("atmospheric muons") can reach the detector (white box) from above. Muons from the lower hemisphere must have been produced in neutrino interactions

arranged at >1 km depth in order to suppress the background from misreconstructed downward moving muons which may mimic upward moving ones (Fig. 17.2).

The identification of extraterrestrial neutrino events faces three sources of backgrounds:

- down-going punch-through muons from cosmic-ray interactions in the atmosphere ("atmospheric muons"). This background can be reduced by going deeper.
- random backgrounds due to photomultiplier (PMT) dark counts, ^{40}K decays (mainly in sea water) or bioluminescence (only water), which impact adversely on event recognition and reconstruction. This background can be mitigated by local coincidences of PMTs.
- neutrinos from cosmic-ray interactions in the atmosphere ("atmospheric neutrinos"). Extraterrestrial neutrinos can be separated from atmospheric neutrinos on a statistical basis (due to their harder energy spectrum). For down-going neutrinos interacting within the detector, atmospheric neutrinos can be largely rejected by vetoing accompanying atmospheric muons from the same shower as the atmospheric neutrino.

Atmospheric neutrinos, of course, have an own scientific value: at medium and high energies they are a well-understood "standard candle" to calibrate the detector, at low energies they allow for investigating neutrino oscillations.

17.2 Neutrino Interactions

The behaviour of the neutrino cross section can be approximated by a linear dependence for $E_\nu < 5$ TeV, for energies larger than 5 TeV by an $E_\nu^{0.4}$ dependence [1]. The absolute value of the cross section at 1 TeV is about 10^{-35} cm^2.

The final state lepton follows the initial neutrino direction with a mean mismatch angle θ decreasing with the square root of the neutrino energy [4]:

$$<\theta> \approx \frac{1.5^\circ}{\sqrt{E_\nu[\text{TeV}]}} \tag{17.1}$$

This on the one hand principally enables source tracing with charged current muon neutrinos, but on the other hand sets a kinematical limit to the ultimate angular resolution. It is worse than for high energy gamma astronomy and particularly worse than for conventional astronomy.

The probability $P_{\nu\to\mu}(E_\nu, E_\mu^{min})$ to produce, in a charged current interaction of a muon neutrino with energy E_ν, a muon reaching the detector with a minimum detectable energy E_μ^{min} depends on the cross section $d\sigma_{\nu N}^{CC}(E_\nu, E_\mu)/dE_\mu$ and the effective muon range R_{eff}, which is defined as the range after which the muon energy has decreased to E_μ^{min} [4]:

$$P_{\nu\to\mu}(E_\nu, E_\mu^{min}) = N_A \int_{E_{\mu,min}}^{E_\nu} dE_\mu \frac{d\sigma_{\nu N}^{CC}(E_\nu, E_\mu)}{dE_\mu} \cdot R_{eff}(E_\mu^{min}, E_\mu) \tag{17.2}$$

with N_A being the Avogadro constant. For water and $E_\mu^{min} \approx 1$ GeV one can approximate [4]

$$P_{\nu\to\mu} = 1.3 \cdot 10^{-6} \cdot E_\nu^{2.2} \text{ for } \mathrm{E}_\nu < 1 \text{ TeV} \tag{17.3}$$

$$= 1.3 \cdot 10^{-6} \cdot E_\nu^{0.8} \text{ for } \mathrm{E}_\nu > 1 \text{ TeV} \tag{17.4}$$

(with E_ν given in TeV). This means, that a telescope can detect a muon neutrino with 1 TeV energy with a probability of about 10^{-6}, if the telescope is on the neutrino's path.

The number of events from a flux Φ_ν recorded by a detector with area A within a time T under a zenith angle ϑ is then given by

$$\frac{N_\mu(E_{\mu,min}, \vartheta)}{AT} = \int_{E_{\mu,min}}^{E_\nu} dE_\nu \Phi_\nu(E_\nu, \vartheta) \cdot P_{\nu\mu}(E_\nu, E_{\mu,min}) \cdot e^{-\sigma_{tot}(E_\nu) N_A Z(\vartheta)} \; . \tag{17.5}$$

Here $Z(\delta)$ is the matter column in the Earth crossed by the neutrino. For sub-TeV energies, absorption in the Earth is negligible and the exponential term $\sim$1 (see Fig. 17.5).

17.3 Principle of Underwater/Ice Neutrino Telescopes

Underwater/ice neutrino telescopes consist of a lattice of photomultipliers (PMs) housed in transparent pressure spheres which are spread over a large volume in oceans, lakes or glacial ice. The PMs record arrival time and amplitude, sometimes even the full waveform, of Cherenkov light emitted by muons or particle cascades.

In most designs the spheres are attached to strings which—in the case of water detectors—are moored at the ground and held vertically by buoys. The typical spacing along a string is 10–25 m, and between strings 60–200 m. The spacing is incomparably large compared to Super-Kamiokande (see chapter C4). This allows covering large volumes but makes the detector practically blind with respect to phenomena below 10 GeV. An exception are planned high-density detectors under water and ice which are tailored to oscillation physics and to the determination of the mass hierarchy of neutrinos [15, 16].

17.3.1 Cherenkov Light

Charged particles moving faster than the speed of light in a medium with index of refraction n, $v \geq c/n$, emit Cherenkov light. The index of refraction depends on the frequency ν of the emitted photons, $n = n(\nu)$. The total amount of released energy is given by

$$-\left(\frac{dE}{dx}\right)_c = \frac{2\pi \cdot \alpha}{c} \cdot \int_{\beta \cdot n(\nu) \geq 1} \left(1 - \frac{1}{\beta^2 \cdot n^2(\nu)}\right) d\nu \quad , \tag{17.6}$$

with α being the fine structure constant and $\beta = v/c$. In the transparency window of water, i.e. for wavelength 400 nm $\leq \lambda \leq$ 700 nm, the index of refraction for water is $n \approx 1.33$, yielding about 400 eV/cm, or ≈200 Cherenkov photons per cm. The spectral distribution of Cherenkov photons is given by

$$\frac{dN}{dx d\lambda} = \frac{2\pi \cdot \alpha}{\lambda^2} \cdot \left(1 - \frac{1}{\beta^2 \cdot n^2}\right) \quad . \tag{17.7}$$

The photons are emitted under an angle Θ_C given by

$$\cos \Theta_C = \frac{1}{\beta \cdot n} \quad . \tag{17.8}$$

For water, $\Theta_C = 41.2°$.

17.3.2 Light Propagation

The propagation of light in water is governed by absorption and scattering. In the first case the photon is lost, in the second case it changes its direction. Multiple scattering effectively delays the propagation of photons. The parameters generally chosen as a measure for these phenomena are [17, 18]:

(a) The absorption length $L_a(\lambda)$—or the absorption coefficient $a(\lambda) = 1/L_a$—with λ being the wavelength. It describes the exponential decrease of the number N of non-absorbed photons as a function of distance r, $N = N_0 \cdot \exp(-r/L_a)$.
(b) The scattering length $L_b(\lambda)$ and scattering coefficient $b(\lambda)$, defined in analogy to $L_a(\lambda)$ and $a(\lambda)$.
(c) The scattering function $\chi(\theta, \lambda)$, i.e. the distribution in scattering angle θ.
(d) Often instead of the "geometrical" scattering length $L_b(\lambda)$, the effective scattering length L_{eff} is used: $L_{eff} = L_b/(1 - \langle\cos\theta\rangle)$ with $\langle\cos\theta\rangle$ being the mean cosine of the scattering angle. L_{eff} "normalizes" scattering lengths for different distributions $\chi(\theta, \lambda)$ of the scattering angle to one with $\langle\cos\theta\rangle = 0$, i.e. L_{eff} is a kind of isotropization length. For $\langle\cos\theta\rangle \sim 0.8-0.95$, as for all media considered here, photon delay effects in media with the same L_{eff} are approximately the same.

Table 17.2 summarizes typical values for Lake Baikal [19, 20], oceans [21, 22] and Antarctic ice [23, 24], each are given for the wavelength of their maximum.

Scattering and absorption in water and ice are determined with artificial light sources. The scattering coefficient in water changes only weakly with wavelength. The dependence on depth over the vertical dimensions of a neutrino telescope in water is small, but parameters may change in time, due to transient water inflows loaded with bio-matter or dust, or due to seasonal changes in water parameters. They must therefore be permanently monitored. In glacial ice at the South Pole, the situation is different. The parameters are constant in time but strongly change with depth (see Sect. 17.6.3).

Strong absorption leads to reduced photon collection, strong scattering deteriorates the time information which is essential for the reconstruction of tracks and showers (see Sects. 17.6 and 17.7).

Table 17.2 Absorption length and effective scattering length for different sites

Site	L_a (m)	L_{eff} (m)
Lake Baikal, 1 km depth	18–22	150–250 (seasonal variations)
Ocean, > 1.5 km depth	40–70 (depends on site and season)	200–300 (depends on site and season)
Polar ice, 1.5–2.0 km depth	∼95 (average)	∼20 (average)
Polar ice, 2.2–2.5 km depth	>100	30–40

17.3.3 Detection of Muon Tracks and Cascades

Neutrinos can interact with target nucleons N through charged current, CC ($\nu_l + N \rightarrow l + X$, with l denoting the charged parter lepton of the neutrino) or neutral current, NC ($\nu_l + N \rightarrow \nu_l + X$) processes. A CC reaction of a ν_μ produces a muon track and a hadronic particle cascade, whereas all NC reactions and CC reactions of ν_e produce particle cascades only. CC interactions of ν_τ can have either signature, depending on the τ decay mode.

In most astrophysical models, neutrinos are expected to be produced through the $\pi/K \rightarrow \mu \rightarrow e$ decay chain, i.e. with a flavour ratio $\nu_e : \nu_\mu : \nu_\tau \approx 1 : 2 : 0$. For sources outside the solar system, neutrino oscillations turn this ratio to $\nu_e : \nu_\mu : \nu_\tau \approx 1 : 1 : 1$ upon arrival on Earth. That means that about 2/3 of the charged current interactions appear as cascades.

Figure 17.3 sketches the two basic detection modes of underwater/ice neutrino telescopes.

17.3.3.1 Muon Tracks

In the muon-track mode, high energy neutrinos are inferred from the Cherenkov cone accompanying muons which enter the detector from below or which strat inside the detector. The upward signature guarantees the neutrino origin of the muon since no other particle can cross the Earth. The effective volume considerably exceeds the actual detector volume due to the large range of muons (about 1 km at 300 GeV and 24 km at 1 PeV [4]).

The muon looses energy via ionization, pair production, bremsstrahlung and photonuclear reactions. The energy loss can be parameterized by [4, 25]

$$-\frac{dE_\mu}{dx} = a + b \cdot E_\mu \quad . \tag{17.9}$$

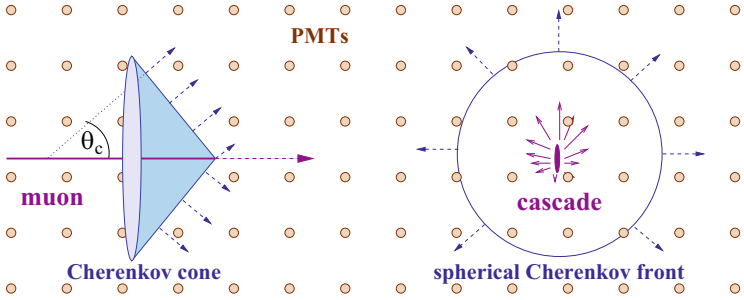


Fig. 17.3 Detection of muon tracks (left) and cascades (right) in underwater detectors

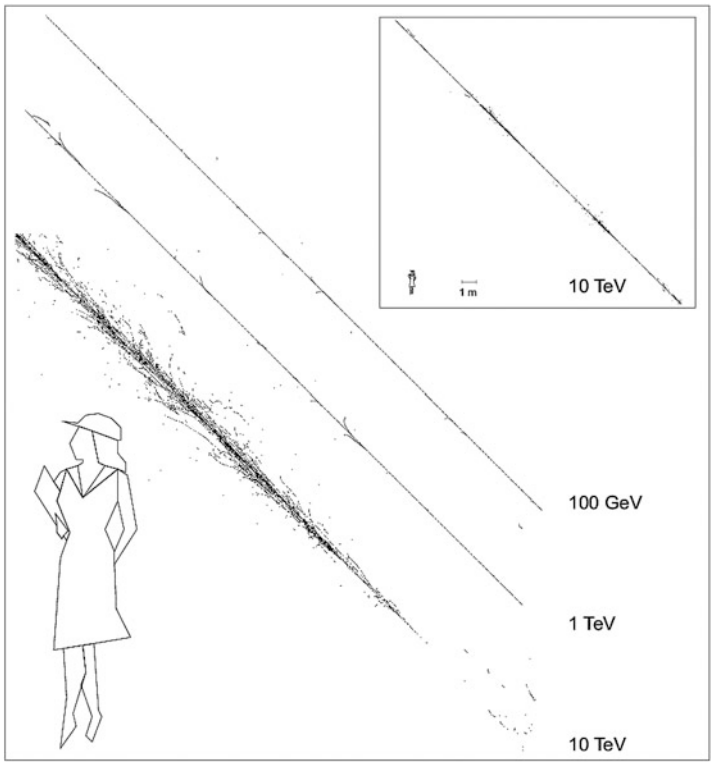


Fig. 17.4 Typical muon tracks and charged secondaries in water above the Cherenkov threshold, for different muon energies (a) 10 TeV (box), and zoomed: 10 TeV, 1 TeV and 100 GeV [26]

For water, the ionization loss is given by $a = 2\,\text{MeV/cm}$, the energy loss from pair production, bremsstrahlung and photonuclear reactions is described by $b = (1.7 + 1.3 + 0.4) \cdot 10^{-6}\,\text{cm}^{-1} = 3.4 \cdot 10^{-6}\,\text{cm}^{-1}$ and rises linearly with energy [25]. Figure 17.4 shows muons tracks with the corresponding secondaries from the last three processes [26]. A detailed description of the muon propagation through matter has to take into account the stochastic character of the individual energy loss processes, which leads to separated cascades of secondaries along the muon track.

Underwater/ice telescopes are optimized for the detection of muon tracks and for energies of a TeV or above, by the following reasons:

(a) The flux of neutrinos from cosmic accelerators is expected to be harder than that of atmospheric neutrinos, yielding a better signal-to-background ratio at higher energies.
(b) Neutrino cross section and muon range increase with energy. The larger the muon range, the larger the effective detector volume.
(c) The mean angle between muon an neutrino decreases with energy like $E^{-0.5}$, resulting in better source tracing and signal-to-background ratio at high energy.

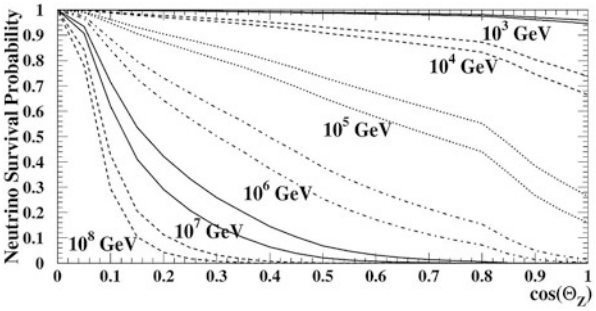


Fig. 17.5 Transmission of the Earth for neutrinos of different energy, as a function of zenith angle [27]

(d) For energies above a TeV, the increasing light emission allows estimating the muon energy with an accuracy of $\sigma(\log E_\mu) \sim 0.3$. By unfolding procedures, a muon energy spectrum can be translated into a neutrino energy spectrum.

Muons which have been generated in the Earth's atmosphere above the detector and punch through the water or ice down to the detector outnumber neutrino-induced upward moving muons by several orders of magnitude (about 10^6 at 1 km depth and 10^4 at 4 km depth) and have to be removed by careful up/down assignment.

At energies above a few hundred TeV, where the Earth is going to become opaque even to neutrinos, neutrino-generated muons arrive preferentially from directions close to the horizon, at EeV energies essentially only from the upper hemisphere (Fig. 17.5). The high energy deposition of muons from PeV-EeV extraterrestrial neutrinos provides a handle to distinguish them—on a statistical basis—from downward going atmospheric muons (those with a spectrum decreasing rather steeply with energy). A different case are down-going muon tracks or cascades starting within the detector. They must be due to neutrino interactions. If the neutrino has been generated in the atmosphere, it will be accompanied in most cases by muons from the same air shower, the higher the energy, the more frequently. Therefore one can apply a veto against accompanying down-going muons and thereby remove most atmospheric neutrinos. This method has been first applied in [8].

17.3.3.2 Cascades

Neutral current interactions and charged current interactions of electron and (most) tau neutrinos do not lead to high energy muons but to electromagnetic or hadronic cascades. Their length increases only like the logarithm of the cascade energy (Fig. 17.6). Cascade events are therefore typically "contained" events.

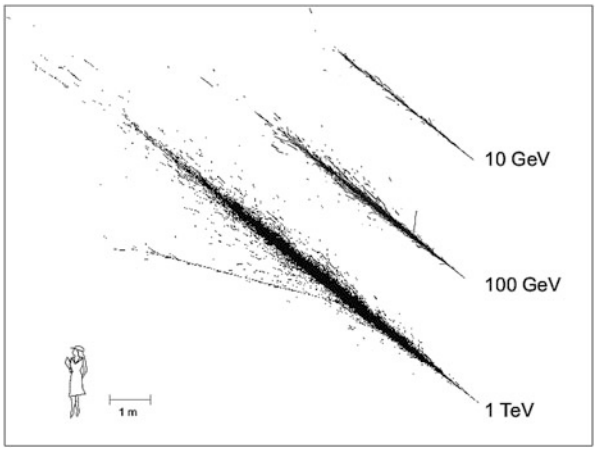


Fig. 17.6 Typical electromagnetic cascades in fresh water for 10 GeV, 100 GeV and 1 TeV [26]

With 5–20 m length in water, and a diameter of the order of 10–20 cm, cascades may be considered as quasi point-like compared to the spacing of the strings along which the PMs are arranged (again with the exception of high-density arrays tailored to oscillation physics). The effective volume for the clear identification of isolated cascades from neutrino interactions is close to the geometrical volume of the detector. For first-generation neutrino telescopes it is therefore much smaller than that for muon detection. However, for kilometer-scale detectors and not too large energies it can reach the same order of magnitude as the latter. The total amount of light is proportional to the energy of the cascade. Since the cascades are "contained", they do not only provide a dE/dx measurement (like muons) but an E-measurement. Therefore, in charged current ν_e and ν_τ interactions, the neutrino energy can be determined with an accuracy of 10–30% (depending on energy and PM spacing). While this is much better than for muons, the directional accuracy is worse since the lever arm for fitting the direction is negligibly small. The background from atmospheric electron neutrinos is much smaller than in the case of extraterrestrial muon neutrinos and atmospheric muon neutrinos. All this taken together, makes the cascade channel particularly interesting for searches for diffuse high-energy excesses of extraterrestrial neutrinos over atmospheric neutrinos.

17.4 Effective Area and Sensitivity

The detection efficiency of a neutrino telescope is quantified by its effective area, e.g., the fictitious area for which the full incoming neutrino flux would be recorded. Fig. 17.7 shows the effective area of the IceCube detector for the detection mode of through-going muons. The increase with E_ν is due to the rise of neutrino cross

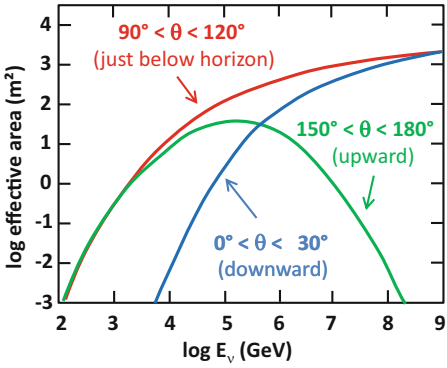


Fig. 17.7 Effective area of the IceCube detector for neutrinos, $A_{eff}(\nu)$, assuming the detection mode of through-going muons. The zenith angle θ is counted 0°/180° for vertically downward/upward moving muons. The effective area is strongly increasing with energy due to increasing neutrino cross section and muon range. The decrease at high energy and large zenith angles is due to the opacity of the Earth to neutrinos with energies above ≈100 TeV. Identification of downward-going neutrinos requires strong cuts against atmospheric muons, hence the cut-off towards low E_ν for $\theta < 30°$

section and muon range, while neutrino absorption in the Earth causes the decrease at large zenith angle θ. Identification of downward-going neutrinos requires strong cuts against atmospheric muons, hence the cut-off towards low E_ν.

Due to the small cross section, the effective area is many orders of magnitude smaller than the geometrical dimension of the detector; a muon neutrino with 1 TeV, e.g., can be detected with a probability of the order 10^{-6} if the telescope is on its path. Note that the detection efficiency for cascades or muons starting within the detector are much smaller since these detection modes do not profit from the potentially large range of muons coming from outside.

Even cubic kilometer neutrino telescopes reach only effective areas between a few square meters and a few hundred square meters, depending on energy. This has to be compared to several ten thousand square meters typical for air Cherenkov telescopes which detect gamma ray-initiated air showers. A ratio 1:1000 (10 m^2:10 000 m^2) may appear desperately small. However, one has to take into account that Cherenkov gamma telescopes can only observe one source at a time, and that their observations are restricted to moon-less, cloud-less nights. Neutrino telescopes observe a full hemisphere, 24 h per day. Therefore, cubic kilometer detectors reach a flux sensitivity similar to that which first-generation Cherenkov gamma telescopes like Whipple and HEGRA [28, 29] had reached for TeV gamma rays, namely $\Phi(>1\,\text{TeV}) \approx 10^{-12}\,\text{cm}^{-2}\,\text{s}^{-1}$.

17.5 Reconstruction

In this section, some relevant aspects of event reconstruction are demonstrated for the case of muons tracks [30, 31]. For cascades, see [32, 33]. The reconstruction procedure for a muon track consists of several consecutive steps:

1. Rejection of noise hits
2. Simple pre-fit procedures providing a first-guess estimate for the following iterative maximum-likelihood reconstructions
3. Maximum-likelihood reconstruction
4. Quality cuts in order to reduce background contaminations and to enrich the sample with signal events. This step is strongly dependent of the actual analysis—diffuse fluxes at high energies, searches for steady point sources, searches for transient sources etc.

An infinitely long muon track can be described by an arbitrary point $\vec{r}_0$ on the track which is passed by the muon at time t_0, with a direction $\vec{p}$ and energy E_0. Photons propagating under the Cherenkov angle θ_c and on a straight path ("direct photons") are expected to arrive at PM i located at $\vec{r}_i$ at a time

$$t_{geo} = t_0 + \frac{\vec{p} \cdot (\vec{r}_i - \vec{r}_0) + d \cdot \tan\theta_c}{c}, \tag{17.10}$$

where d is the closest distance between PM i and the track, and c the vacuum speed of light. The time residual t_{res} is given by the difference between the measured hit time t_{hit} and the hit time expected for a direct photon t_{geo}:

$$t_{res} = t_{hit} - t_{geo}. \tag{17.11}$$

Schematic distributions for time residuals are given in Fig. 17.8. An unavoidable symmetric contribution in the range of a nanosecond comes from the PM/electronics time jitter, σ_t. An admixture of noise hits to the true hits from a muon track adds a flat pedestal contribution like shown in top right of Fig. 17.8. Electromagnetic cascades along the track lead to a tail towards larger (and only larger) time residuals (bottom left). Scattering of photons which propagate in water loaded with bio-matter and dust or in ice can lead to an even stronger delay of the arrival time (bottom right). These residuals must be properly implemented in the probability density function for the arrival times.

The simplest likelihood function is based exclusively on the measured arrival times. It is the product of all N_{hit} probability density functions p_i to observe, for a given value of track parameters $\{a\}$, photons at times t_i at the location of the hit PMs:

$$L_{time} = \prod_{i=1}^{Nhit} p(t_{res,i}|\{a\}) \tag{17.12}$$

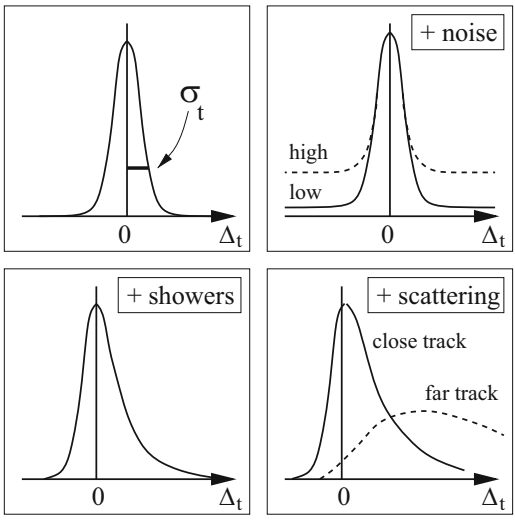


Fig. 17.8 Schematic distributions of arrival times for different cases (see text)

More complicated likelihoods include the probability of hit PMs to be hit and of non-hit PMs to be not hit, or the amplitudes of hit PMs. Instead for referring only to the arrival time of the first photon for a given track hypothesis, and the amplitude for a given energy hypothesis, one may also refer to the full waveform from multiple photons hitting the PM. For efficient background suppression, the likelihood may also incorporate information about the zenith angular dependence of background and signal (Bayesian probability). The reconstruction procedure finds the best track hypothesis by maximizing the likelihood.

17.6 First Generation Neutrino Telescopes

The development of the field was pioneered by the project DUMAND (Deep Underwater Muon And Neutrino Detection Array) close to Hawaii [34]. First activities started in 1975. With the final goal of a cubic kilometre array, the envisaged first step was a configuration with 216 optical modules at 9 strings, 30 km offshore the Big Island of Hawaii, at a depth of 4.8 km. A test string with 7 optical modules (OMs) was deployed in 1987 from a ship, took data at different depths for several hours and measured the depth dependence of the muon flux [35]. A shore cable for a stationary array was laid in 1993 and a first string with 24 OMs deployed. It failed due to water leakages. Financial and technical difficulties led to the official termination of the project in 1996. Therefore the eventual breakthrough and proof of principle came from the other pioneering experiment located in Lake Baikal. See for the history of neutrino telescopes [36].

17.6.1 The Baikal Neutrino Telescope NT200

The Baikal Neutrino Telescope NT200 was installed in the Southern part of Lake Baikal [37]. The distance to shore is 3.6 km, the depth of the lake at this location is 1366 m, the depth of the detector about 1.1 km.

The BAIKAL collaboration was not only the first to deploy three strings (as necessary for full spatial reconstruction), but also reported the first atmospheric neutrinos detected underwater [38, 39] (see also Fig. 17.9, right).

NT200 was an array of 192 optical modules (OMs), completed in April 1998. It is sketched in Fig. 17.9, left. The OMs were attached to eight strings carried by an umbrella-like frame consisting of 7 arms each 21.5 m in length. The strings were anchored by weights at the bottom and held in a vertical position by buoys at various depths. The geometrical dimensions of the configuration were 72 m (height) and 43 m (diameter). Detectors in Lake Baikal are deployed (or hauled up for repairs) within 6–7 weeks in February/April, when the lake is covered with a thick ice layer providing an ideal, stable working platform. They are connected to shore by several cables which allow operation over the full year.

The time calibration of NT200 was done with several nitrogen lasers, one sending short light pulses via optical fibres of equal length to each individual OM pair (top

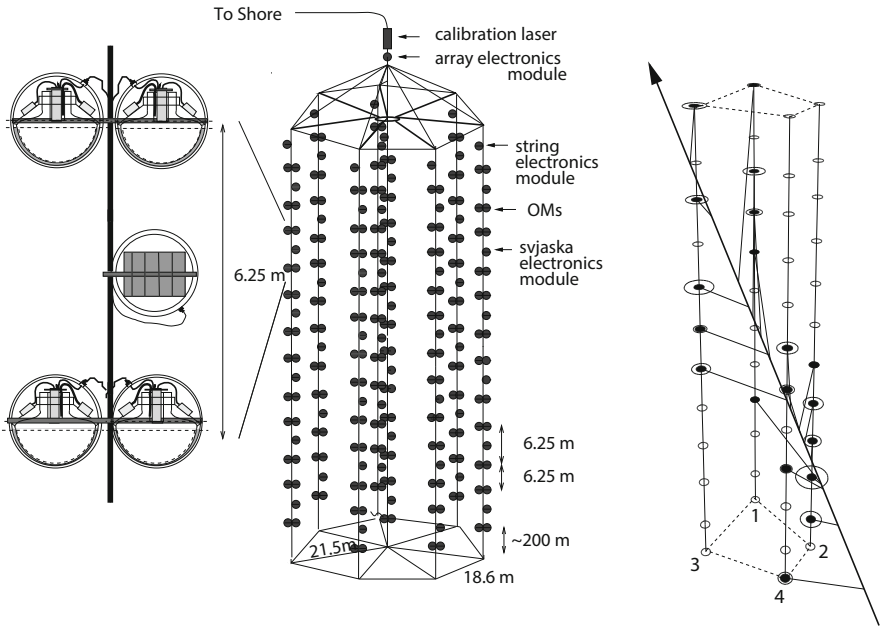


Fig. 17.9 Left: The Baikal Neutrino Telescope NT200. Right: One of the first upward moving muons from a neutrino interaction recorded with the 4-string stage of the Lake Baikal detector in 1996 [40]. The muon fires 19 channels

of Fig. 17.9), the other, the light pulses of the other laser below the array (not shown in the figure) propagate to the OMs through the water.

The OMs consisted of a pressure glass housing equipped with a QUASAR-370 phototube and were grouped pair-wise along a string. In order to suppress accidental hits from dark noise (∼30 kHz) and bio-luminescence (typically 50 kHz but seasonally raising up to hundreds of kHz), the two PMs of a pair were switched in coincidence, defining a *channel*, with only ∼0.25 kHz noise rate. The basic cell of NT200 consisted of a *svjaska* (Russian for "bundle"), comprising two OM pairs and an electronics module which was responsible for time and amplitude conversion and slow control functions (Fig. 17.9, left). A majority trigger was formed if $\geq m$ channels were fired within a time window of 500 ns (this is about twice the time a relativistic particle needed to cross the NT200 array), with m typically set to 4. Trigger and inter-string synchronization electronics were housed in an array electronics module at the top of the umbrella frame. This is less than 100 m away from the OMs, allowing for easy nanosecond synchronization over copper cable.

Figure 17.10 shows the phototube and the full OM [41]. The QUASAR-370 consisted of an electro-optical preamplifier followed by a conventional PM (type UGON). In this hybrid scheme, photoelectrons from a large hemispherical cathode (K_2CsSb) with $>2\pi$ viewing angle are accelerated by 25 kV to a fast, high gain scintillator which is placed near the centre of the glass bulb. The light from the scintillator is read out by the small conventional PM. One photoelectron emerging from the hemispherical photocathode yields typically 20 photoelectrons in the conventional PM. This high multiplication factor results in an excellent single electron resolution of 70%, a small time jitter (2 ns) and a small sensitivity to the Earth's magnetic field. The OM contains the QUASAR, the HV supply for the small PM (2 kV) and the large tube (25 kV) and a LED. The signal from the last dynode and the anode is read out via two penetrators, the two other penetrators pass the signal driving the calibration LED and the low voltages for the HV system and the preamplifiers. The optical contact between QUASAR bulb and glass housing is

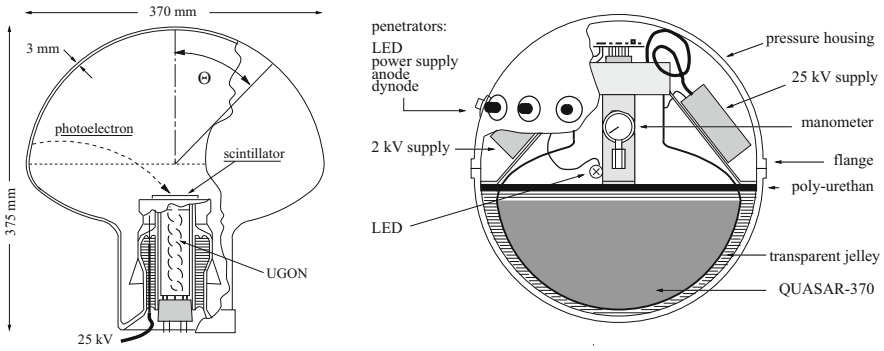


Fig. 17.10 Left: The QUASAR phototube. Right: a full Baikal optical module [41]

made by liquid glycerine sealed with a layer of polyurethane, in later versions with a silicon gel.

Due to the small lever arm, the angular resolution of NT200 for muon tracks was only 3–4°. On the other hand, the small spacing of modules led to a comparably low energy threshold for muon detection of ∼15 GeV. The total number of upward muon events collected over 5 years was only about 400, due not only to the small dimensions of the array, but also to its unstable operation. Still, NT200 could compete for some time with the much larger AMANDA, by searching for high energy cascades *below* NT200, surveying a volume about ten times as large as NT200 itself [42].

17.6.2 AMANDA

Rather than water, AMANDA (Antarctic Muon And Neutrino Detection Array) was using the 3 km thick ice layer at the South Pole as target and detection medium [43, 44]. AMANDA was (actually still *is*, although switched off) located some hundred meters away from the Amundsen–Scott station which provides the necessary infrastructure. Holes of 60 cm diameter were drilled with pressurized hot water, and strings with OMs were deployed in the column of molten water and frozen into the ice. South Pole installation operations are performed in the Antarctic summer, November to February, when temperatures rise to up to −25 °C. For the rest of the time, two operators (of a winter-over crew of 30–40 persons in total) maintain the detector, connected to the outside world via satellite communication.

Figure 17.11, left, shows the configuration of AMANDA. A first shallow test array with 80 OMs at 4 strings (not shown in the figure) was deployed in the Antarctic season 1993/1994, at depths between 800 and 1000 m [45]. It turned out that the effective scattering length L_{eff} was desperately small, 40 cm at 830 m depth, but increased with depth (80 cm at 970 m depth). The scattering was due to remnant bubbles and made track reconstruction impossible. The tendency of scattering decreasing with depth, as well as results from ice core analyses at other places in Antarctica, suggested that bubbles should disappear below 1300 m. This expectation was confirmed with a second 4-string array which was deployed in 1995/1996. The effect of bubbles disappeared, with the remaining scattering being mostly due to dust (see Fig. 17.11, right). The scattering length averaged over 1500–2000 m depth is $L_{eff} \approx 20$ m, still considerably worse than for water but sufficient for track reconstruction [30, 46]. The array was upgraded stepwise, completed in January 2000 and eventually comprised 19 strings with a total of 677 OM, most of them at depth between 1500 and 2000 m.

Figure 17.11, right, gives absorption and scattering coefficient as a function of depth and wavelength for glacial ice at the South Pole. The variations with depth are due to (a) bubble remnants at shallow depth leading to very strong scattering, (b) dust and other scattering and absorbing material transported in varying climate epochs to Antarctica. The depth dependence complicates the evaluation of the

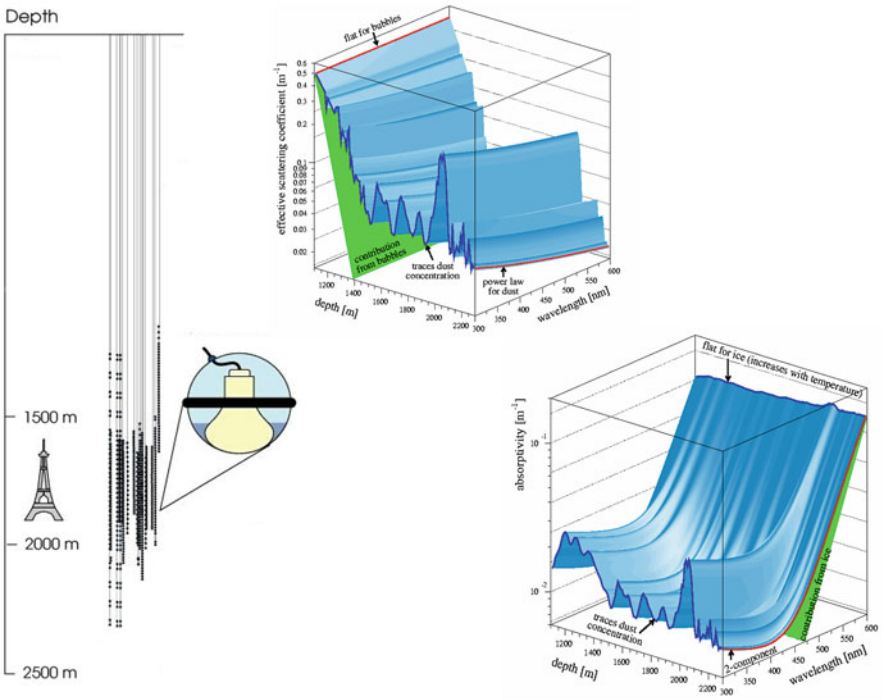


Fig. 17.11 Left: The AMANDA configuration. Three of the 19 strings have been sparsely equipped towards larger and smaller depth in order to explore ice properties, one string got stuck during deployment at too shallow depth and was not used in analyses. Right: scattering coefficient (top) and absorption coefficient (bottom) as a function of depth and wavelength

experimental data. Furthermore, the strong scattering leads to strong delays in photon propagation, resulting in worse angular resolution of deep ice detectors compared to water. On the other hand, the large absorption length, with a cut-off below 300 nm instead at 350–400 nm (water), results in better photon collection than in water. The quality of the ice improves substantially below a major dust layer at 2000–2100 m, with a value for the scattering length about twice as large as for the shallower region above 2000 m.

The short distance between OM and surface electronics allowed for a unique technical solution: the analogue PM anode signals were not digitized in the depth, but driven over 2 km cable to surface. This requires a large output signal of the PM, a specification met by the 8-inch R5912-2 from Hamamatsu with 14 dynodes and an internal amplification of 10^9. The first ten strings used coaxial (string 1–4) and twisted pair (string 6–10) cables for both HV supply and signal transmission, for the last 9 strings the anode signal was fed to an LED, and the light signal transmitted via optical fibre to surface. Naturally, the electrical signal transmission suffered from strong dispersion, widening the anode signal to several 100 ns. However, applying an amplitude correction to time flags from a constant fraction discriminator, a time

jitter of 5–7 ns was achieved. Given the strong smearing of photon arrival times due to light scattering in ice, this jitter appeared to be acceptable. For optical signals, dispersion was negligible. An event was defined by a majority trigger formed in the surface counting house, requesting $\geq$8 hits within a sliding window of 2 µs.

Time calibration of the AMANDA array was performed with a YAG laser at surface (wavelengths >450 nm), sending short pulses via optical fibres of well defined length to each OM. This laser system was also used to measure the delay of optical pulses propagating between strings and to determine the ice properties as well as the inter-string distances. A nitrogen laser (337 nm) at 1850 m depth, halogen lamps (350 and 380 nm) and LED beacons (450 nm) extended the information about ice properties across a large range of wavelengths (see Fig. 17.11, right). The measured time delays were fitted and the resulting parameterizations implemented in the probability density functions for the residual times t_{res}.

One big advantage compared to underwater detectors is the small PM noise rate, about 1 kHz in an 8-inch PM, compared to 20–40 kHz due to K^{40} decays and bioluminescence in lakes and oceans. The contribution of noise hits to the true hits from a particle interaction is therefore small and makes hit cleaning procedures much easier than in water.

The angular resolution of AMANDA for muon tracks was 2–2.5°, with an energy threshold of ≈50 GeV. Although better than for Lake Baikal (3–4°), it was much worse than for ANTARES (<0.5°, see below). This is the result of the strong light scattering which deteriorates the original information contained in the Cherenkov cone. The effect is even worse for cascades, where the angular resolution achieved with algorithms of that time was only ≈25 deg (compared to 5–8° in ANTARES).

In 2008, AMANDA had established a series of record upper limits, e.g. for diffuse extraterrestrial neutrino fluxes using muon as well as cascade searches, for the flux of relativistic magnetic monopoles or for neutrinos from point sources (see for a review [47]). The final AMANDA point source analysis was based on 6595 neutrinos collected in the years 2000–2006 [48]. AMANDA was switched off in 2009.

17.6.3 Mediterranean Projects: ANTARES

Mediterranean efforts to build an underwater neutrino telescope are related to three locations:

(a) a site close to Pylos at the Peloponnesus, with available depths ranging from 3.5 to 5 km for distances to shore of 30–50 km,
(b) a site close to Capo Passero, Sicily, at a depth of 3.5 km and 70 km, distance to shore,
(c) a site close to Toulon, at a depth of 2.5 km and 40 km distance to shore.

All of these sites are considered locations for a future distributed infrastructure of a total volume of few cubic kilometres. All sites have physics and infrastructural

pros and cons. For instance, large depth is a challenge for long-term ocean technology and bears corresponding risks, but has convincing physics advantages: less background from punch-through muons from above, less bio-luminescence, less sedimentation.

Historically, the first Mediterranean project was NESTOR [49] off the Greek coast. It was conceived as a tower-like structure with 12 floors, 300 m in height and 32 m in diameter. A prototype hexagonal floor with 14 PMs (15-inch Hamamatsu) was deployed in 2004 and took data for a few weeks. The project is terminated meanwhile. NEMO, close to Sicily [50], focused on technology development and feasibility studies for a cubic kilometer array. The basic unit of NEMO have been towers composed by a sequence of floors. The floors consist of rigid horizontal structures, 15 m long, each equipped with four 10-inch PMs. The floors are tilted against each other and form a three-dimensional structure.

In the following, ANTARES [51, 52] is described, being the one of the three projects which made it to a full telescope of AMANDA-size.

Figure 17.12 shows a schematic view of the detector. It consists of 12 strings, each anchored at the seabed and kept vertical by a top buoy. The minimum distance

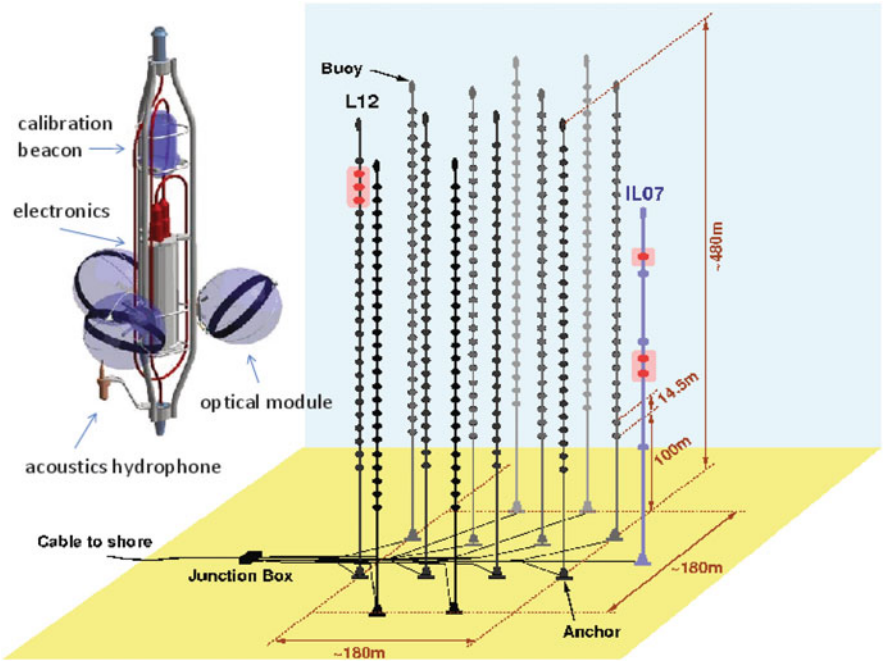


Fig. 17.12 Right: Schematic view of the ANTARES detector, with 12 detector lines L1–L12, and an extra-line with environmental equipment (IL07). L12 and ILO7 carry test equipment for basic tests toward acoustic neutrino detection. Left: A storey with three optical modules and the metallic cylinder housing the Local Control Module (LCM). Every fifth storey carries a LED beacon (above the LCM) and a hydrophone (bottom left) for acoustic triangulation[52]

between the strings is 60 m. Each string is composed of 25 storeys. A storey is equipped with three 10-inch PMs Hamamatsu R7091-20 housed in 13-inch glass spheres. The PMs are oriented at 45° with respect to the vertical. A mu-metal grid reduces the influence of the Earth magnetic field. The storeys are spaced by 14.5 m, the lowest being located about 100 m above seabed. The storeys are connected by an electro-optical cable, including 21 optical fibres for digital communications [53, 54].

From a functional point of view, each string is divided into five sectors, each containing five storeys. A storey is controlled by a Local Control Module (LCM) which maintains the data communication between its sector and the shore. A String Controller Module (SCM), located at the basis of each string, interfaces the string to the rest of the detector. The string cables are led to a junction box to which the shore cable is connected.

The signals from the PMTs are digitized by an Analogue Ring Sampler (ARS). The ARS produces "hits" by time-stamping the PMT signal and by integrating the PMT anode current over a programmable time interval (25–80 ns). The time stamp is provided by the local clock of the LCM. The master clock signal of 20 MHz is generated at shore and distributed through optical fibres to the LCM clocks. Sub-nanosecond precision is achieved by a time-to-voltage converter which allows interpolation between two subsequent clock pulses. The output voltage is digitized with an 8-bit ADC. The maximally achievable time resolution is therefore $1/(20\,\text{MHz} \times 256) \sim 0.2\,\text{ns}$.

The timing calibration is performed with calibration pulses between shore clock and LMC clocks, and with LED beacons which fire both the ARS (electrically) and the PMT (optically) and correct for the varying PMT transit time. The position calibration is particularly import since the string positions change due to water current. It is performed with compasses and tiltmeters along the strings, and with a acoustic triangulation system based on transmitters at the bottom of the strings and hydrophones along the strings. The relative positions of the OMs can be determined with an accuracy of a few centimetres.

The Monte Carlo angular resolution for muons is 0.2° at 10 TeV. At low energies the neutrino tracing is limited by the angle between muon and neutrino, 0.7° at 1 TeV and 1.8° at 100 GeV (median mismatch angle for those muons triggering the detector [55]).

Naturally, the angular resolution for cascades is worse than for muons. Simple reconstruction algorithms give 10° median mismatch angle above 5 TeV, however, with proper quality cuts, values below 4° can be achieved, with 20–40% passing rates for signals [33].

ANTARES is operated in its full configuration since 2008 and is planned to continue data taking until the follow-up project KM3NeT has surpassed ANTARES w.r.t. to its sensitivity, i.e. at least through 2018.

17.7 Second Generation Neutrino Telescopes

17.7.1 IceCube

IceCube [56] is the successor of AMANDA. It consists of 5160 digital optical modules (DOMs) installed on 86 strings at depths between 1450 and 2450 m in the Antarctic ice [57], and 320 DOMs installed in IceTop [58], detectors in pairs on the ice surface directly above the strings (see Fig. 17.13). AMANDA was integrated into IceCube as a low-energy sub-detector, but later was replaced by DeepCore, a high density, six-string sub-array at large depths (i.e. in best ice) at the centre of IceCube. The energy threshold is about 100 GeV for the full IceCube array and about 10 GeV for DeepCore.

The thermal power of hot-water drill factory is increased to 5 MW, compared to 2 MW for AMANDA, reducing the average time to drill a 60 cm hole to 2450 m depth down to ≈35 h. The subsequent installation of a string with 60 DOMs requires

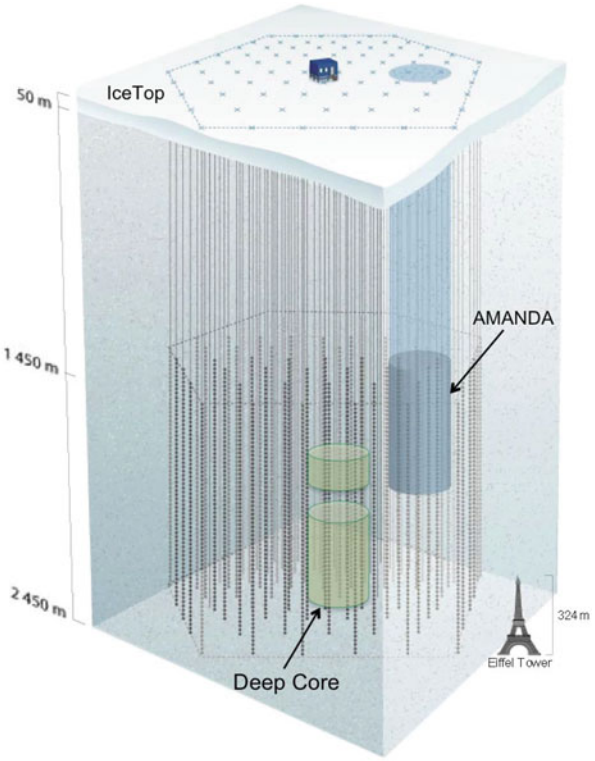


Fig. 17.13 Schematic view of the IceCube Neutrino Observatory. Since 2009, AMANDA is replaced by DeepCore, a nested low-threshold array. At the surface are the air shower detector IceTop and the IceCube counting house

typically 12 h. A record number of 20 strings was deployed in the season 2009/2010. The detector was completed in December 2010.

The components are not accessible after refreezing of the holes. Therefore—as for AMANDA—the architecture has to avoid single-point failures in the ice. A string carries 60 DOMs, with 30 twisted copper pair cables providing power and communication. Two sensors are operated on the same wire pair. Neighbouring DOMs are connected to enable fast local coincidence triggering in the ice [57].

A schematic view of a DOM is shown in Fig. 17.14. A 10-inch PMT Hamamatsu R7081-02 is embedded in a 13-inch glass pressure sphere [59]. A mu-metal grid reduces the influence of the Earth's magnetic field. The programmable high voltage is generated inside the DOM. The average PMT gain is set to 10^7. Signals are digitized by a fast analogue transient waveform recorder (ATWD, 3.3 ns sampling) and by a FADC (25 ns sampling). The PM signal is amplified by 3 different gains to extend the dynamic range of the ATWD to 16 bits, resulting in a linear dynamic range of 400 photoelectrons in 15 ns; the dynamic range integrated over 2 μs is about 5000 photoelectrons.

The digital electronic on the main board are based on a field-programmable gate array (FPGA). It communicate with the surface electronics, new programs can be downloaded. The LEDs on the flasher board emit calibration pulses at 405 nm which can be adjusted over a wide range up to $\sim 10^{11}$ photons.

All digitized PM pulses are sent to the surface. The full waveform, however, is only sent for pulses from local (neighbour or next to neighbour) coincidences in order to apply data compression for isolated hits which are mostly noise pulses. All DOMs have precise quartz oscillators providing local clock signals, which are synchronized every few seconds to a central GPS clock. The time resolution is about

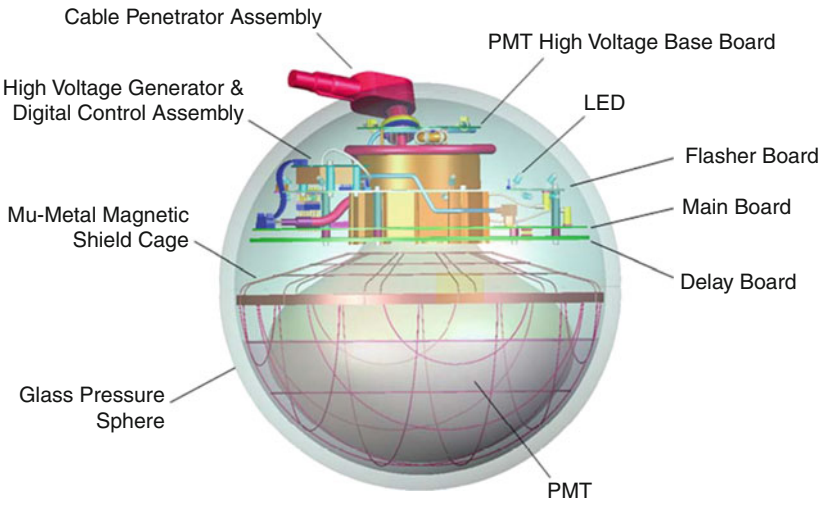


Fig. 17.14 Schematic view of an IceCube digital optical module

2 ns. The noise rate for DOMs in the deep ice is ∼540 Hz, if a deadtime of 250 μs is applied only ∼280 Hz. The very low noise rates are critical for the detection of the low-energy neutrino emission associated with a supernova collapse (see below).

At the surface, 8 custom PCI cards per string provide power, communication and time calibration. Subsequent processors sort and buffer hits until the array trigger and event builder process is completed [61]. The architecture allows deadtime free operation. The design raw data rate of the full array is of the order of 100 GB/day which are written to tapes. Online processing in a computer farm allows extraction of interesting event classes, like all upgoing muon candidates, high-energy events, IceTop/IceCube coincidences, cascade events, events from the direction of the muon or events in coincidence with Gamma Ray Bursts (GRB). The filtered data stream (∼20 GB/day) is then transmitted via satellite to the Northern hemisphere.

The muon angular resolution is about 1° for 1 TeV tracks and below <0.5° for energies of 10 TeV and higher. The very good ice below 2100 m has a particular potential for improved resolution. This will be even more important for the angular reconstruction of cascades. The presently achieved angular resolution for cascades is only 25°, much worse than for water, with the inferiority being mainly due to light scattering in ice.

IceCube is the only detector which can be permanently operated together with a surface air shower array, IceTop [58]. It consists of tanks filled with ice, each instrumented with 2 DOMs. The comparison of air shower directions measured with IceTop and directions of muons from these showers in IceCube allows an angular calibration of IceCube (absolute pointing and angular resolution). IceTop can measure the spectrum air showers up to primary particle energies of $\sim 10^{18}$ eV. Combination of IceTop information (reflecting dominantly the electron component of the air shower) and IceCube information (muons from the hadronic component) allows estimating the mass range of the primary particle.

Last but not least, IceCube allows for another mode of operation which is essentially only possible in ice: the detection of burst neutrinos from a supernova [60]. The low dark counting rate of PMs (∼280 Hz, see above) allows detecting of the feeble increase of the summed count rates of all PMs during several seconds, which would be produced by millions of MeV neutrino interactions from a supernova burst. IceCube records the counting rate of all PMs in millisecond steps. A supernova in the centre of the Galaxy would be detected with extremely high confidence and the onset of the pulse could be measured in unprecedented detail. Even a 1987A-type supernova in the Large Magellanic Cloud would result in a 5σ effect and be sufficient to provide a trigger to the SuperNova Early Warning System, SNEWS [62].

The following figures show displays of some events recorded with IceCube. Figure 17.15, left, is a typical muon track crossing the detector from below. The event on the right side is a cascade event, actually the fully contained cascade event with the highest energy recorded, about 2 PeV. The analysis employed containment conditions and an atmospheric muon veto for suppression of down-going atmospheric neutrinos ("High-Energy Starting Event" analysis, HESE). The HESE events cannot be explained by atmospheric neutrinos and misidentified

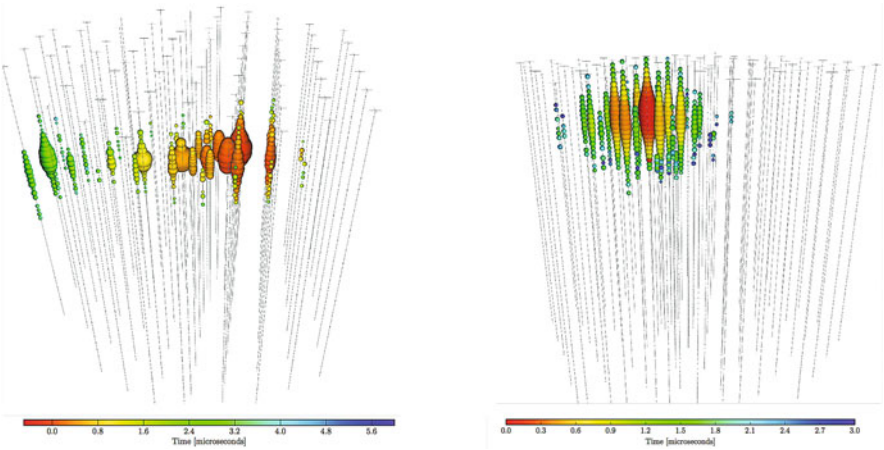


Fig. 17.15 Left: A through-going upward muon track. Right: The highest-energy cascade event detected (status 2018) with IceCube, with ≈2 PeV energy released in the detector [68]. The size of the symbols reflect the recorded amount of light, the color indicates the signal timing (red: early; green: late), see the scale at the bottom

atmospheric muons alone: with 6 years of data, the excess has a significance of $> 7\sigma$, i.e. a flux of extraterrestrial neutrinos could be safely confirmed.

Also, events with through-going muons show a corresponding excess of cosmic origin [69]—see the display of the highest-energy track-like neutrino event in Fig. 17.16.

In its final configuration, IceCube takes data since spring 2011, with a duty cycle of more than 99%. It collects almost 10^5 clean neutrino events per year, with nearly 99.9% of them being of atmospheric origin. The failure rate of DOMs is only about one per year, out of more than 5000.

17.7.2 KM3NeT

KM3NeT has two main, independent objectives: (a) the discovery and subsequent observation of high-energy cosmic neutrino sources and (b) precise oscillation measurements and the determination of the mass hierarchy of neutrinos [63, 64]. For these purposes the KM3NeT Collaboration plans to build an infrastructure distributed over three sites: off-shore Toulon (France), Capo Passero (Sicily, Italy) and Pylos (Peloponnese, Greece). In a configuration to be realized until 2021/2022, KM3NeT will consist of three so-called building blocks ("KM3NeT Phase-2").

A building block comprises 115 strings, each string with 18 optical modules.Two building blocks will be sparsely configured to fully explore the IceCube signal with a comparable instrumented volume, different methodology, improved resolution and complementary field of view, including the Galactic plane. These two blocks will

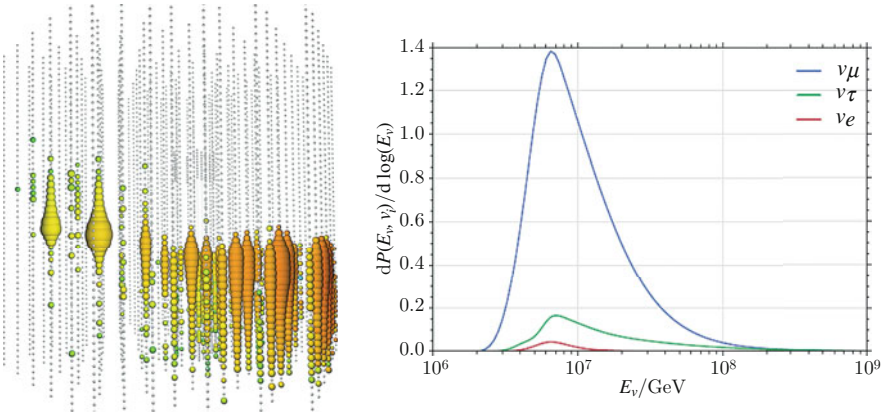


Fig. 17.16 Left: Event view of the PeV track-like event recorded by IceCube on June 11, 2014. Color code like in the previous figure. Note that the scaling is non-linear and a doubling in sphere size corresponds to one hundred times the measured charge. This event deposited an energy of 2.6 ± 0.3 PeV in the detector volume. Right: Probability distribution of primary neutrino energies that could result in the observed multi-PeV track-like event, assuming an E_ν^{-2} spectrum. The total probabilities for the different flavors are 87.7, 10.9 and 1.4% for ν_μ, ν_e and ν_τ, respectively. The most probable energy of the primary neutrino is between 8 and 9 PeV

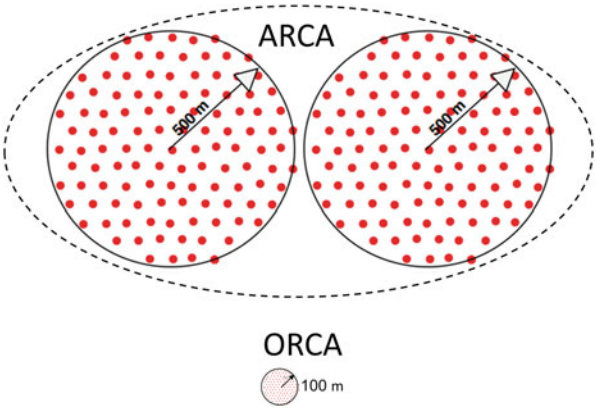


Fig. 17.17 The two incarnations of KM3NeT. The two ARCA blocks (top) have diameters of 1 km and a height of about 600 m and focus to high-energy neutrino astronomy. ORCA (bottom) is a shrinked version of ARCA with only 200 m diameter and 100 m height. Both ARCA and ORCA have 115 strings with 18 optical modules (OMs) per string

be deployed at the Capo Passero site and are referred to as ARCA: Astroparticle Research with Cosmics in the Abyss. The third building block will be densely configured to precisely measure atmospheric neutrino oscillations. This block, being deployed at the Toulon site, is referred to as ORCA: Oscillation Research with Cosmics in the Abyss (see Fig. 17.17).

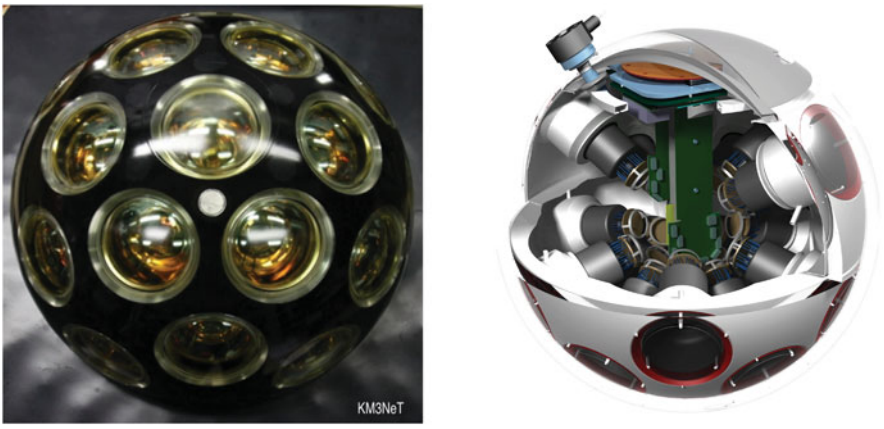


Fig. 17.18 View and a cross-sectional drawing of a KM3NeT-DOM with its 31 small PMs inside [64]

A novel concept has been chosen for the KM3NeT optical module: The 43 cm glass spheres of the DOMs will be equipped with 31 PMs of 7.5 cm diameter, with the following advantages: (a) The overall photocathode area exceeds that of a 25 cm PM by more than a factor three; (b) The individual readout of the PMs results in a very good separation between one- and two-photoelectron signals which is essential for online data filtering; (c) some directional information is provided. This technical design has been validated with in situ prototypes. A view and a cross-sectional drawing of the DOM are shown at the top of Fig. 17.18.

Rather than digitizing the full waveform (like for the one large PM per DOM in IceCube), for each of the analogue pulses from 31 small PMs which pass a preset threshold, the time of the leading edge and the time over threshold are digitized (referred to as a *hit*). Each hit corresponds to 6 Bytes of data (1 B for PM address, 4 B for time and 1 B for time over threshold, with the least significant bit of the time information corresponding to 1 ns). All hits are sent to shore (all-data-to-shore concept). The total rate for a single building block with its 64,170 PMTs amounts to about 25 Gb/s which are sent via optical fibers to shore. To limit the number of fibres, wavelength multiplexing is used.

At shore, the physics events are filtered from the background. To maintain all available information for the offline analysis, each event contains a snapshot of all the data during that event. The filtered data (with a rate reduced by a factor of about 10^5 with respect to the data arriving at shore), are stored at disks.

KM3NeT-ARCA is conceived as the European counterpart to IceCube and will preferentially observe the Southern instead of the Northern hemisphere, including the Galactic Centre [63]. With a fully equipped ARCA, IceCube's cosmic neutrino flux could be detected with high-significance within 1 year of operation. In practise the detector will be deployed in stages allowing to reach the 1 year sensitivity of two clusters much before the second cluster is fully installed.

ORCA will continue along the venue opened by IceCube-DeepCore and perform precision measurements of neutrino oscillations. In particular, it could determine the neutrino mass hierarchy with at least 3σ significance after 3 years of operation.

17.7.3 GVD

Based on the long-term experience with the NT200 detector and on extended prototype tests, the Baikal Collaboration has started the stepwise installation of a kilometer-scale array in Lake Baikal, the Giant Volume Detector, GVD [65, 66].

The optical modules of Baikal-GVD are equipped with 10-inch PMs of the type Hamamatsu R7081-100, with a quantum efficiency of ≈35%. The OMs are mounted on vertical strings, fixed to the bottom with anchors. Each eight strings form a cluster, with 36 OMs per string, i.e. 288 OMs per cluster. Each cluster is a full functional detector which is capable of detecting a physical event both in standalone mode and as part of the full-scale array. The first phase of GVD (GVD-1) is planned to be be completed by 2021, with eight clusters carrying $\approx 2.3 \times 10^3$ OMs in total and a volume of 0.3–0.4 km^3. Figure 17.19 shows a schematic view of GVD-1. In a second phase, Baikal-GVD is conceived to be extended to an array of about 10^4 OMs with an instrumented volume of 1–2 km^3.

The OMs are vertically spaced by 15 m, with the lowest OM at a depth of 1275 m (about 100 m above the bottom of the lake) and the top OM at 750 m below the lake surface. The seven strings of a cluster are arranged at a radius of 60 m around a central string. The distances between the centers of the clusters are 300 m.

A string is composed of three *sections*, each comprising 12 OMs with analog outputs. A *Central section Module (CM)* converts the analog signals into a digital code, using a 12-bit ADC with a sampling frequency of 200 MHz. Coincidences of signals from any pairs of neighbouring OMs are used as a local trigger of the section (signal request), with average frequencies of the section request signals in the range

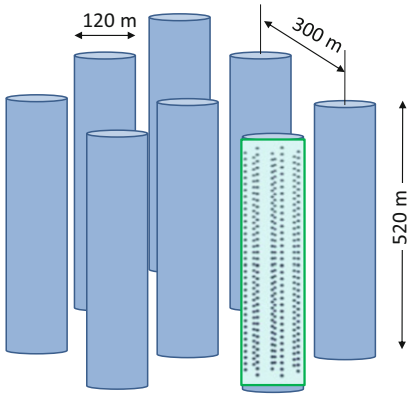


Fig. 17.19 Schematic view of phase-1 of Baikal-GVD, consisting of 8 clusters, each with 120 m diameter and 520 m height. A cluster consists of eight strings with 36 optical modules along each string

of 2–10 Hz, dependent on signal thresholds and level of water luminescence. The request signals from three sections are combined in the *Control Module* of the string (CoM) and transferred to the cluster Cluster DAQ center where a global trigger is formed. The Cluster DAQ center is arranged close to the water surface at a depth of 25 m and connected to the Shore DAQ Center by hybrid electro-optical cable.

Calibration is performed by LEDs and lasers. LEDs installed in each OM provide amplitude and time calibration of the OMs, separate underwater modules equipped with LEDs are used for time calibration between sections. A high-power laser is used arranged between clusters ensures calibration of the cluster as a whole and calibration between neighboured clusters. The coordinates of the optical modules are determined using an acoustic positioning. Each cluster has its own acoustic positioning system, with four acoustic modems per string, the lowest at the bottom of the string, the highest 538 m higher. The transit time between acoustic sources at the lake bed and the acoustic modems gives the coordinates of the acoustic modems with an accuracy of $\approx$2 cm.

17.7.4 IceCube-Gen2

The progress from IceCube will be limited by the modest numbers of cosmic neutrinos measured, even in a cubic kilometer array. In [67] a vision for the next-generation IceCube neutrino observatory is presented. At its heart is an expanded array of optical modules with a volume of 7–10 km^3. This high-energy array will mainly address the 100 TeV to 100 PeV scale. For point sources, it will have five times better sensitivity than IceCube, and the rate for events at energies above a few hundred TeV will be ten times higher than for IceCube. It has the potential to deliver first GZK neutrinos, of anti-electron neutrinos produced via the Glashow resonance, and of PeV tau neutrinos, where both particle showers associated with the production and decay of the tau are observed ("double bang events").

Another possible component of IceCube-Gen2 is the PINGU sub-array. It targets—similar to ORCA—precision measurements of the atmospheric oscillation parameters and the determination of the neutrino mass hierarchy. The facility's reach would further be enhanced by exploiting the air-shower measurement and vetoing capabilities of an extended surface array. Moreover, a radio array ("ARA", for Askarian Radio Array, see below) will achieve improved sensitivity to neutrinos in the $10^{16}-10^{20}$ eV energy range, including GZK neutrinos.

17.8 Physics Results: A 2018 Snapshot

The 2018 status of the field is dominantly defined by the IceCube results. ANTARES significantly contributes to searches for neutrinos from the Southern hemisphere and the central parts of the Galaxy. These are the main results obtained over the last 5 years:

- Both IceCube and ANTARES have measured the flux of "conventional" atmospheric neutrinos from π and K decay up to a few hundred TeV and found it in agreement with predictions [70, 71]. Tight upper limits have been set for the flux of "prompt" atmospheric neutrinos from charm and bottom decays.
- At energies below 50 GeV, the oscillation of atmospheric neutrinos passing through the Earth has been observed both by IceCube and ANTARES. The IceCube constraints on the neutrino mixing parameters are meanwhile as tight as those derived from accelerator experiments [72].
- In 2013, IceCube has detected a diffuse flux of astrophysical neutrinos with a very high confidence (meanwhile larger than 7σ). This observation can be considered a real breakthrough, 53 years after the first ideas on underwater neutrino detectors have been proposed [73].
- ANTARES and IceCube have jointly analysed their data to identify a neutrino excess from the Galactic Plane and can exclude that more than 8.5% of the observed diffuse astrophysical flux comes from the Galactic plane [74].
- No steady neutrino point sources could be identified, neither using 8 years of IceCube data, with 497,000 upward muons from neutrino interactions, nor with ANTARES data. The derived limits on point source fluxes are a fantastic factor 3000 below those obtained in 2000 with AMANDA data [75].
- Also, various analyses where many sources belonging to a certain source class are "stacked" did not yield significant excesses. For instance, latest IceCube results exclude that more 6% of the observed diffuse astrophysical muon neutrino flux could come from blazars (active galaxies with their jet pointing to the Earth) [76]. Blazars have been considered since long as top-candidate neutrino sources. The same applies to neutrinos from Gamma Ray Bursts (GRB). IceCube could exclude at more than 90% confidence those models which assume that GRBs are the dominant source of the measured cosmic-ray flux at highest energies [77].
- An alert issued by IceCube on September 22, 2017, led to the first coincident observation of a high-energy energy neutrino with X-ray, gamma-ray and optical information. These electromagnetic follow-up observations identified a blazar named TXS 0506+056 in its active state as the likely source of the neutrino. IceCube examined its archival data in the direction of TXS 0606+056 and found an additional 3.5σ evidence for a flare of 13 neutrinos starting at the end of 2014 and lasting about 4 months. This is considered the first compelling evidence for flaring source of neutrinos [78, 79].
- No neutrinos from cosmic-ray interactions with the 3K-microwave background radiation could yet be identified. Their observation will need multi-km^3 detectors like IceCube-Gen2 or even radio detectors as discussed in the next Sect. [80].

- Record limits have been derived for neutrino fluxes from dark matter annihilations in the Earth, the Sun or the Galactic halo and for the flux of magnetic monopoles (which, if at relativistic velocity, could be identified via their high light emission) and to the coupling of hypothetical sterile neutrinos to normal neutrino states (see [81] for a review of results on particle physics with IceCube).

17.9 Technologies for Extremely High Energies

The technologies described in this section are tailored to signals which propagate with km-scale attenuation. Consequently, they allow for the observation of much larger volumes than those typical for optical neutrino telescopes. 100 km^3 scale detectors are necessary, for instance, to record more than just a few GZK neutrinos, with a typical energy range of 100 PeV to 10 EeV.

17.9.1 Detection via Air Showers

At energies above 10^{17} eV, large extensive air shower arrays like the Pierre Auger detector in Argentina [82] or the Telescope Array in Utah/USA [83] are seeking for horizontal air showers due to neutrino interactions deep in the atmosphere (showers induced by charged cosmic rays start on top of the atmosphere). Figure 17.20 explains the principle. AUGER consists of an array of water tanks spanning an area

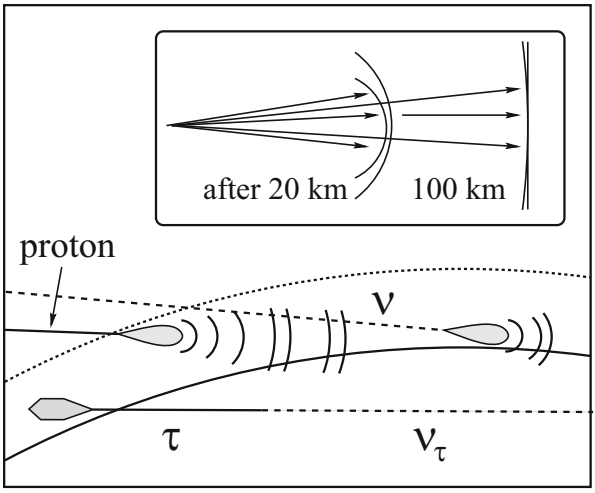


Fig. 17.20 Detection of particles or fluorescence light emitted by horizontal or upward directed air showers from neutrino interactions

of 3000 km^2 and recording the Cherenkov light of air-shower particles crossing the tanks. It is combined with telescopes looking for the atmospheric fluorescence light from air showers (see chapter on cosmic ray detectors). The optimum sensitivity window for this method is at 1–100 EeV, the effective detector mass is up to 20 Gigatons. An even better sensitivity might be obtained for tau neutrinos, ν_τ, scratching the Earth and interacting close to the array [84, 85]. The charged τ lepton produced in the interaction can escape the rock around the array, in contrast to electrons, and in contrast to muons it decays after a short path into hadrons. If this decay happens above the array or in the field of view of the fluorescence telescopes, the decay cascade can be recorded. Provided the experimental pattern allows clear identification, the acceptance for this kind of signals can be large. For the optimal energy scale of EeV, the present differential single-flavor limit (2017) is about $2 \times 10^{-8} E_\nu^{-2}$ GeV^{-1} cm^{-2} s^{-1} sr^{-1} [86].

A variation of this idea is to search for tau lepton cascades which are produced by horizontal PeV neutrinos hitting a mountain and then decay in a valley between target mountain and an "observer" mountain [87].

17.9.2 Radio Detection

Electromagnetic cascades generated by high energy neutrino interactions in ice or salt emit coherent Cherenkov radiation at radio frequencies. The effect was predicted in 1962 [88] and confirmed by measurements at accelerators [89, 90]. Electrons are swept into the developing shower, which acquires an electric net charge from the added shell electrons. This charge propagates like a relativistic pancake of 1 cm thickness and 10 cm diameter. Each particle emits Cherenkov radiation, with the total signal being the convolution of the overlapping Cherenkov cones. For wavelengths larger than the cascade diameter, coherence is observed and the signal rises proportional to E_ν^2, making the method attractive for high energy cascades. The bipolar radio pulse has a width of 1–2 ns. In ice, attenuation lengths of up to a kilometer are observed, depending on the frequency band and the ice temperature. Thus, for energies above a few ten PeV, radio detection becomes competitive or superior to optical detection (with its attenuation length of $\sim$100 m) [91].

A prototype Cherenkov radio detector called RICE was operated at the South Pole, with 20 receivers and emitters buried at depths between 120 and 300 m. From the non-observation of very large pulses, limits on the diffuse flux of neutrinos with E$>$100 PeV and on the flux of relativistic magnetic monopoles have been derived [92].

Three groups are working towards detectors with 100–300 km^3 active volume: The Askarian Radio Array (ARA [93]) at the South Pole, the Antarctic Ross Iceshelf Antenna Neutrino Array (ARIANNA [94]) on the Antarctic Ross ice shelf—both in the phase of tests with engineering arrays—and the Greenland Neutrino Observatory

(GNO [96]) which is conceived to be deployed near the USA Summit Station in Greenland.

The current ARA proposal [93] envisages and array of 37 stations, each consisting of 16 antennas, buried up to 200 m depth below the firn ice. The stations are spaced by 2 km. As of 2018, five of them are deployed. ARIANNA [94] will observe the 570 m thick ice covering the Ross Sea. The smooth ice-seawater interface reflects radio waves; therefore ARIANNA might have a better sensitivity for downward moving and horizontal neutrinos. However, the ice is warmer than at the South Pole, reducing the attenuation length for GHz radio waves from 800–900 m (South Pole) to about 400 m (ice shelf). ARIANNA antennas face downward and are arranged just below the ice surface, with about thousand antennas for the ultimate array, spread over an area of $\approx$1000 km^2. One can reasonably expect that only one of these two projects can be funded in its full size.

ANITA (Antarctic Impulsive Transient Array [95] is an array of radio antennas which has been flown at a balloon on an Antarctic circumpolar path in 2006 and 2008/2009 (see Fig. 17.21).

From 35 km altitude it searches for radio pulses from neutrino interactions in the thick ice cover and monitored, with a threshold in the EeV range and a volume of the order of 10^6 Gigatons. This corresponds to a much larger volume than that of ARA and ARIANNA and can be achieved only for the price of an energy threshold about two orders of magnitude above that of ARA and ARIANNA. With its dual-polarization horn antennas it scanned the ice out to 650 km away. Neutrino signals would be vertically polarized, while background signals from down-going cosmic-ray induced air showers are preferentially horizontally polarized. Signals pointing to known or suspected areas of human activity are rejected. The ANITA 90% C.L. integral flux limit on a pure E^{-2} spectrum, integrating over $10^{18}-10^{23.5}$ eV, is $E^2 \times 1.3 \cdot 10^{-7}$ GeV cm^{-2} s^{-1} sr^{-1}, presently (2018) the most stringent limit on the GZK neutrino flux.

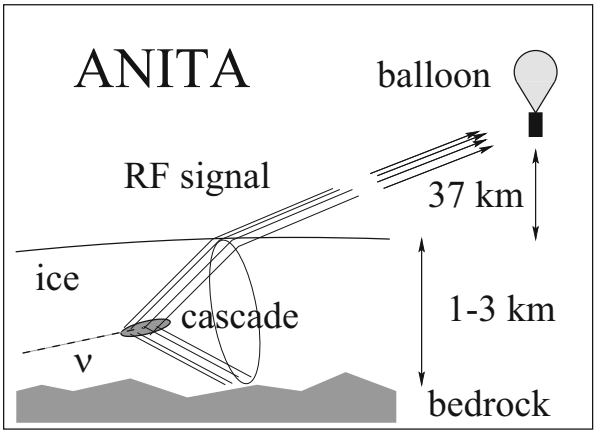


Fig. 17.21 Principle of the ANITA balloon experiment

Even higher energies are are addressed when searching for radio emission from particle cascades induced by neutrinos or cosmic rays skimming the moon surface. An example is the GLUE project (Goldstone Ultra-high Energy Neutrino Experiment [97]) which used two NASA antennas and reached maximum sensitivity at several ZeV (1 ZeV = 1000 EeV). With the same method, the NUMOON experiment at the Westerbork Radio Telescope searched for extremely energetic neutrinos [98], and the LUNASKA experiment which uses the Parkes and ATCA radio telescopes [99]. LUNASKA stands for "Lunar Ultra-high Neutrino Astrophysics with the SKA", indicating the final purpose: to use the Square Kilometer Array SKA to perfrom a lunar neutrino search.

17.9.3 Acoustic Detection

Production of pressure waves by fast particles passing through liquids was predicted in 1957 [100] and experimentally proven with high intensity proton beams two decades later [101]. A high energy cascade deposits energy into the medium via ionization losses which is immediately converted into heat. The effect is a fast expansion, generating a bipolar acoustic pulse with a width of a few 10 µs in water or ice (Fig. 17.22). Transversely to the pencil-like cascade, the radiation propagates within a disk of about 10 m thickness (the length of the cascade) into the medium. The signal power peaks at 20 kHz where the attenuation length of sea water is a few kilometres, compared to 50 m for light. The threshold of this method is however very high, in the several-EeV range. Acoustic detection was also considered an option for ice, where the signal itself is higher and ambient noise is lower than in water. A test array, SPATS (South Pole Acoustic Test Setup), has been deployed at the South Pole in order to determine attenuation length and ambient noise [102]. Another test configuration has been deployed together with the ANTARES detector (see Fig. 17.22). Tests are also performed close to Sicily, close to Scotland and in Lake Baikal. Another project has been using a very large hydrophone array of the US Navy, close to the Bahamas [109]. The existing array of hydrophones spans an area of 250 km and has good sensitivity at 1–500 kHz and can trigger on events above 100 EeV with a tolerable false rate.

17.9.4 Hybrid Arrays

Best signal identification would be obtained by combining signatures from two of the three methods, optical, radio and acoustic [110]. Naturally, radio detection does not work in water. The threshold for acoustic detection is so high that coincidences from a 100 km^3 acoustic array and a 1 km^3 optical array would be rare and a true hybrid approach not promising. The hybrid principle may be applicable at the South Pole, since the overlap between optical and radio methods (threshold for radio ∼10–

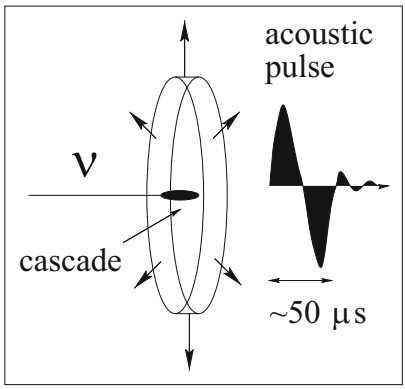


Fig. 17.22 Acoustic emission of a particle cascade

100 PeV) is significant. A nested hybrid array with optical-radio coincidences is therefore be conceivable and is actually part of the IceCube-Gen2 proposal (see previous section).

Overviews on acoustic and radio detection can be found in the proceedings of the workshops on "Acoustic and Radio EeV Neutrino Detection Activities" (ARENA) [103–108].

References

1. T.K. Gaisser, F. Halzen and T. Stanev, Phys. Rep. 258 (1995) 173.
2. J.G. Learned and K. Mannheim, Ann. Rev. Nucl. Part. Sci. 50 (2000) 679.
3. U.F. Katz and C. Spiering, Prog. in Part. Nucl. Phys. 67 (2012) 651 and arXiv:1111.0507.
4. R. Engel, T. Gaisser and E. Resconi, *Cosmic Rays and Particle Physics*, Cambridge University Press 2016.
5. T. Gaisser and A. Karle (eds.),*Neutrino Astronomy*, World Scientific 2017.
6. V.S. Berezinsky, G.T. Zatsepin, Phys. Lett. B28(1969) 423.
7. K. Greisen, Phys. Rev. Lett. 16 (1966) 748; G.T. Zatsepin and A.A. Kuzmin, J. Exp. Theor. Phys. Lett. 4 (1966) 78.
8. M. Aartsen et al. (IceCube Coll.), Science 342 (2013) 2342856.
9. A.B. McDonald et al., Rev. Sci. Instrum. 75 (2004) 293, and arXiv:0311343.
10. M.A. Markov, Proc. ICHEP, Rochester (1960) 578.
11. U. Katz and C. Spiering in C. Patrignani et al. (Particle Data Group), Chin. Phys. C. 40 (2016) 10001.
12. L. Anchordoqui, F. Halzen, Annals Phys. 321 (2006) 2660.
13. M. Ahlert, C. de los Heros and K. Helbing, review to appear in Europ. Phys. Journ. C.
14. M. G. Aartsen et al. (IceCube Coll.), arXiv:1707.07081.
15. P. Coyle for the KM3NeT Coll., J. Phys. Conf. Ser. 888 (2017) no.1, 012024 and arXiv:1701.01382.
16. M.G. Aartsen et al. (IceCube Coll.), J. Phys. G44 (2017) 054006 and arXiv:1707.02671.
17. N.G. Jerlov, Marine Optics, Elsevier Oceanography Series 5 (1976).
18. H. Bradner and G. Blackington, Appl. Opt. 23 (1984) 1009.
19. V. Balkanov et al., Appl. Opt. 33 (1999) 6818.
20. V. Balkanov et al., Nucl. Instrum. Meth. A298 (2003) 231.

21. J.A. Aguliar et al., Astropart. Phys. 23 (2005) 131, and arXiv:astro-ph/0412126.
22. G. Riccobene et al., Astropart. Phys. 27 (2007) 1, and arXiv:astro-ph/0603701 (see also for further references therein).
23. M. Ackermann et al., J. Geophys. Res. 111 (2006) D13203.
24. J. Lundberg et al., Nucl. Instrum. Meth. A581 (2007) 619.
25. W. Lohmann et al. CERN Yellow Report 85–03.
26. C.H.W. Wiebusch, PhD thesis, preprint PITHA 95/37.
27. I. Albuquerque, J. Lamoureux and G.F. Smoot, Astrophys. J. Suppt. 114 (2002) 195 and arXiv:hep-ph/0109177.
28. T.C. Weekes et al., Astrophys. J.343 (1989) 379.
29. D. Petry et al., J. Astron. Astrophys. 311 (1996) L13.
30. J. Ahrens et al., Nucl. Instrum. Meth. A524 (2004) 169.
31. Y. Becherini for the Antares coll., Proc. 30th Int. Cosmic Ray Conf. Merida 2007, and arXiv:0710.5355 (see also references therein).
32. M. Ackermann et al., Astropart. Phys. 22 (2004) 127, and arXiv:astro-ph/0405218.
33. B. Hartmann, PhD thesis, Erlangen 2008, see arXiv:astro-ph/06060697.
34. A. Roberts, Rev. Mod. Phys. 64 (1992) 259.
35. E. Babson et al.,Phys. Rev. D42 (1990) 3613.
36. C. Spiering Eur. Phys. J. H37 (2012) 515 and arXiv:1207.4952.
37. I.A. Belolaptikov et al., Astropart. Phys. 7 (1997) 263.
38. CERN Courier, Sept. 1996, p.24.
39. C. Spiering for the Baikal Coll., Prog. Part. Nucl. Phys. 40 (1998) 391.
40. R.V. Balkanov et al., Astropart. Phys. 12 (1999) 75, and arXiv:astro-ph/9705244.
41. R. Bagduev et al., Nucl. Instrum. Meth. A420 (1999) 138.
42. V. Aynutdinov et al., Astropart. Phys. 25 (2006) 140, and arXiv:astro-ph/0508675.
43. E. Andres et al., Astropart. Phys. 13 (2000) 1.
44. E. Andres et al., Nature 410 (2001) 441.
45. P. Askebjer et al., Science 267 (1995) 1147.
46. M. Ackermann et al., J. Geophys. Res. 111 (2006) D13203.
47. T. DeYoung, Journ. of Physics Conf. Series 136 (2008) 042058.
48. R. Abbasi et al. Phys. Rev. D79 (2009) 062001.
49. G. Aggouras et al., Nucl. Instr. Meth. A552 (2005) 420.
50. E. Migneco et al., Nucl. Instr. Meth. A588 (2008) 111.
51. ANTARES homepage, http://antares.in2p3.fr
52. M. Ageron et al. (ANTARES Coll.) Nucl. Instr. Meth. A656 (2011) 11 and arXiv:1104.1607.
53. J. Aguilar, Astropart. Phys. 26 (2006) 314.
54. M. Ageron et al., Astropart. Phys. 31 (2009) 277. and arXiv:0812.2095.
55. T. Montaruli, J. of Modern Physics A, arXiv:0810.3933.
56. IceCube homepage, http://icecube.wisc.edu.
57. M.G. Aartsen et al. (IceCube Coll.) JINST 12 (2017) P03012 and arXiv:1612.05093.
58. R. Abbasi et al. (IceCube Coll.), Nucl. Instr. Meth. A 700 (2013) 188 and arXiv:1207.6326.
59. R. Abbasi et al. (IceCube Coll.), Nucl. Instr. Meth. A 618 (2010) 139 and arXiv:1002.2442.
60. R. Abbasi et al. (IceCube Coll.) Astron. Astrophys. 535 (2011) A109 and arXiv:1108.0171.
61. R. Abbasi et al. (IceCube Coll.), Nucl. Instr. Meth. A 601 (2009) 294 and arXiv:0810.4930.
62. SNEWS: P. Antonioli et al., New Journ. Phys. 6 (2004) 114 and arXiv:astro-ph/0406214.
63. KM3NeT homepage: http://www.km3net.org.
64. S. Adrian Martinez et al. (KM3NeT Coll.) J.Phys. G43 (2016) no.8, 084001 and arXiv:1601.07459.
65. http://baikalweb.jinr.ru, including a full english project description.
66. V. Avronin et al. (Baikal Coll.), Nucl. Instr. Meth. A742 (2014) 82. Nucl. Instr. Meth. A602 (2009) 227, and arXiv:0811.1110.
67. M.G. Aartsen et al. (IceCube Coll.), arXiv:1412.5106.
68. M.G. Aartsen et al. (IceCube Coll.), Phys. Rev. Lett. 113 (2014) 101101.
69. M.G. Aartsen et al. (IceCube Coll.) Astrophys. J. 833 (2016) no.1, 3 and arXiv:1607.08006.

70. R. Abbasi et al. (IceCube Coll.) Phys. Rev. D83 (2011) 012001 and arXiv:1010.3980.
71. S. Adrian-Martinez et al. (Antares Coll.) Eur. Phys. J. C73 (2013) 2606 and arXiv:1306.1599.
72. M.G. Aartsen et al. (IceCube Coll.) Phys. Rev. Lett. 120/no.7 (2018) and arXiv:1707.07081.
73. M.G. Aartsen et al. (IceCube Coll.) Science 342 (2013) 1242856 and arXiv:1311.5238.
74. A. Albert et al. (Antares and IceCube Coll.) Astrophys. J. 868 (2018) L20 and arXiv:1808.03531.
75. M.G. Aartsen et al. (IceCube Coll.) arXiv:1811.07979.
76. M.G. Aartsen et al. (IceCube Coll.) Astrophys. J. 835/no.1 (2017) 45 and arXiv:1611.0374.
77. M.G. Aartsen et al. (IceCube Coll.) Astrophys. J. 843/no.2 (2017) 112. and arXiv:1702.06862.
78. M.G. Aartsen et al. (IceCube, Fermi-LAT, MAGIC, AGILE, ASAS-SN, HAWC, H.E.S.S., INTEGRAL, Kanata, Kiso, Kapteyn, Liverpool Telescope, Subaru, Swift NuSTAR, VERITAS and VLA/17B-403 Collaborations) Science 361/no.6389 (2018) eaat1378 and arXiv:1807.08816.
79. M.G. Aartsen et al. (IceCube Coll.) Science 361/no.6389 (2018) 147 and arXiv:1807.08794.
80. M.G. Aartsen et al. (IceCube Coll.) Phys. Rev. D98 (2018) 062003 and arXiv:1807.01820
81. M. Ahlers, K. Helbig and C. Perez de los Heros, Eur. Phys. J.C 78 (2018) 924 and arXiv:1806.05695.
82. A. Aab et al., Nucl. Instr. Meth. A798 (2015) 172.
83. H. Tokuno et al., Nucl. Instr. Meth. A676 (2012) 54.
84. Letessier Selvon, Proc. AIP Conf. 566 (2001) 157.
85. D. Fargion, Astrophys. Journ. 570 (2002) 909.
86. A. Aab et al., Phys.Rev. D91 (2015) no.9, 092008.
87. G.W.S. Hou and M.A. Huang, astro-ph/0204145.
88. G.A. Askaryan, Sov.Phys. JETP 14 (1962) 441.
89. D. Saltzberg et al., Phys. Rev. Lett. 86 (2001) 2802.
90. P. Gorham et al., Phys. Rev. Lett. 99 (2007) 171101.
91. B. Price, Astropart. Phys. 5 (1996) 43.
92. I. Kravtchenko et al., Phys. Rev. D73 (2006) 082002.
93. P. Allison et al. (ARA Coll.), Phys.Rev. D93 (2016) no. 8, 082003 and arXiv:1507.08991.
94. S. Barwick et al. (ARIANNA Coll.) Astropart. Phys. 90 (2017) 50 and arXiv:1612.04473.
95. P. Gorham et al. (ANITA Coll.) Phys. Rev. Lett. 103 (2009) 051103.
96. S.A. Wissel et al., PoS (ICRC 2015) 1150. Astroparticle Physics 90 (2017) 50 and arXiv:1612.04473.
97. P. Gorham et al., Phys. Rev. Lett. 93 (2004) 0041101.
98. O. Scholten et al., in [Proc. 2008 ARENA Workshop, Rome 2008, Nucl. Instr. Meth. 2009, ed. A. Capone] and arXiv:0810.3426.
99. J. D. Bray, Phys. Rev. D91 (2015) no.6, 063002 and arXiv:1502.03313.
100. G.A. Askaryan, Sov. Journ. Atom. Energy 3 (1957) 921.
101. see e.g. J.G. Learned, Phys. Rev. D1, 19 (1979) 3293.
102. R. Abbasi et al. (IceCube Coll.) Astropart.Phys. 34 (2011) 382 and arXiv:1004.1694; R. Abbasi et al. (IceCube Coll.), Astropart.Phys. 35 (2012) 312 and arXiv:1103.1216.
103. Proc. Int. Workshop on Acoustic and Radio EeV Neutrino detection Activities (ARENA) DESY, Zeuthen 2005, World Scientific 2006, eds. R. Nahnhauer and S. Boeser.
104. Proc. 2006 ARENA Workshop, Univ. Northumbria 2006, Journ. of Phys., Conf. Series 81 (2007), eds. L. Thomson and S. Danaher.
105. L. Thompson, Nucl. Instr. Meth. 558 (2008) 155.
106. Proc. 2008 ARENA Workshop, Rome 2008, Nucl. Instr. Meth. 2009, ed. A. Capone.
107. Proc. 2012 ARENA Workshop, eds. R. Lahmann, Th. Eberl, K. Graf, C. James, T. Huege, T. Karg and R. Nahnhauer, AIP Conference Collection Volume 1535, Erlangen 2012.
108. Proc. 2017 ARENA Workshop, eds. S. Buitnik, J.R. H"orandel, S. de Jong, R. Lahmann, R. Nahnhauer, O. Scholten, EPJ Web Conf. 135 (2017).
109. N. Lehtinen et al., Astropart. Phys. 17 (2002) 279 and astro-ph/010433.
110. D. Besson et al., in [Proc. 2008 ARENA Workshop, Rome 2008, Nucl. Instr. Meth. 2009, ed. A. Capone] and arXiv:0811.2100.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 18
Spaceborne Experiments

Roberto Battiston

18.1 Introduction: Particle Physics from Ground to Space

The Universe is the ultimate laboratory to understand the laws of nature. Under the action of the fundamental forces, lasting infinitesimal times or billion of years, matter and energy reach most extreme conditions. Using sophisticated instruments capable to select the signals reaching us from the depths of space and of time, we are able to extract information otherwise not obtainable with the most sophisticated ground based experiments. The results of these observations deeply influence the way we today look at the Universe and try to understand it.

During hundreds of thousands of years we have observed the sky only using our eyes, accessing in this way only the very small part of the electromagnetic radiation which is able to traverse the atmosphere, the visible light. The first telescope observations by Galileo in 1609, which dramatically changed our understanding of the solar system, yet were based only on the visible part of the electromagnetic spectrum. Only during the second half of the twentieth century we started to access wider parts of the spectrum. After the end of the war, using the new radar related technology, the scientists developed the radio telescopes to record the first radio images of the galaxy. But only in the 60's, with the advent of the first man made satellites, we began to access the much wider e.m. spectrum, including infrared, UV, X-ray and γ-ray radiation.

A similar situation happened with the charged cosmic radiation. Cosmic Rays, discovered by Hess in 1912 [1] using electrometers operated on atmospheric balloons, for about 40 years were the subject of very intense studies. The discovery of a realm of new particles using CR experiments, gave birth to particle physics

R. Battiston (✉)
Dipartimento di Fisica, Università di Trento, Povo, Italy
e-mail: roberto.battiston@pg.infn.it; roberto.battiston@unitn.it

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_18

and to high energy physics, successfully performed, since the 50's, at particle accelerators. However, the study of the cosmic radiation performed within the atmosphere, deals only with secondary particles. The primary radiation can only be studied with stratospheric balloons or using satellites. In 1958 Van Allen [2] and collaborators studied for the first time the charged cosmic radiation trapped around the Earth, and, since, the measurement of Cosmic Rays from space has become an important tool for the study of the Universe.

A third, more recent example is the discovery of ravitational waves (GW) [3], one hundred years after the prediction of Einstein [4]. Direct observation of gravitational waves opened a new era in astrophysics, adding to the spectrum of electromagnetic radiation the new messenger represented by GW. Since the pioneering attempts of Weber [5] in the 60's, using resonating bars, the GW community has developed in the 90's a network of $O(1)$ km arms, ground based interferometers to search for GWs in the frequency range 10 Hz to 100s of Hz [6],[7]. The detected signals confirm the prediction of General Relativity but also validate the sensitivity of the interferometer technologies. GW are expected much more abundantly in the frequency range $O(0.001)$ Hz to a $O(1)$ Hz. This range can be studied with a $5 \cdot 10^6$ km arm, space based interferometer, as the proposed ESA/NASA LISA mission. The successful LISA technology demonstrator, the ESA lead LISA-Pathfinder (LISA-PF) [8] flown in 2016, opened the way for the LISA[9] adoption, to be developed and implemented during the 20's to start operating at the end of the 20's or a the beginning of the 30's.

During the last century, particle detectors developed on ground have been adapted or designed to be used on stratospheric balloons and on space born experiments. Space, however, is a hostile environment and launching a payload is a very expensive endeavour. For these reasons, the design and the testing of a spaceborn detector requires particular care. In this chapter we deal with this topic.

We begin discussing the properties of the space environment from the upper atmosphere, the transition from the atmosphere to the magnetosphere and from the magnetosphere to the deep interplanetary space.

We then address the requirements for hardness and survivability of space born instrumentation.

We subsequently turn to the issue of manufacturing of hardware to be operated in space, with particular care to the issue of the space qualification tests.

We will also discuss modern spaceborne high energy radiation detectors, mainly from the point of view of the design characteristics related to the operation in space. We will make no attempt to cover the historical development or to cover low energy radiation instrumentation, in particular X-ray space borne detectors.

18.2 The Space Environment

18.2.1 The Neutral Component

Although there are some notable exceptions, a good fraction of scientific satellites which observe the different kinds of radiations emitted by the universe operate on LEO (Low Earth Orbit), namely between 200 and 2000 km from the Earth surface. Below 200 km the atmospheric drag dramatically reduces the lifetime of satellites, above 700 km the radiation environment, due to the Van Allen belts, becomes more and more hostile.

When operating close to the lower limit of LEO orbits, the external surfaces of the payloads are affected by the heat produced by the upper atmosphere drag and by the corrosion due to the presence of highly reactive elements such as atomic oxygen. Above ∼600 km drag is sufficiently weak not to influence anymore the lifetime of most satellites.

Altitudes below ∼600 km are within the Earth's *thermosphere*, the region of the atmosphere where the absorption of the solar UV radiation induces a fast rate of temperature increase with the altitude. At ∼200–250 km the temperature of the tenuous residual atmosphere reaches a limiting value, the *exospheric temperature* ranging from ∼600–1200 K over a typical solar cycle. The thermosphere temperature can also quickly change during the geomagnetic activity.

Atomic oxygen is the main atmospheric constituent from ∼ 200–600 km, since it is lighter than molecular nitrogen and oxygen. Figure 18.1 shows the altitude profiles of atomic oxygen for different solar activities. Atomic oxygen plays an important role in defining the properties of LEO space environment. Since this form of oxygen is highly reactive, surfaces covered with thin organic films, advanced composites or thin metallized layers can be damaged [10]. Kapton, for example, erodes at a rate of approximately 2.8 μm for every 10^{24} atoms/m^2 of atomic oxygen fluence [11], with the fluence during a time interval t being defined as $F_0 = \rho_N v t$, ρ_N being the number density of atomic oxygen and v the satellite velocity. Chemical reaction involving atomic oxygen can in turn produce excited atomic states emitting significant amount of e.m. radiation, creating effects such as the *shuttle glow* which are interfering with optical instrumentation.

18.2.2 The Thermal Environment

From a thermal point of view a spacecraft orbiting around the Earth is exposed to various heat sources; direct sunlight, sunlight reflected off the Earth or other planets (*albedo*) and infrared radiation emitted by the planet atmosphere or surface. The spacecraft in turn loses energy by radiation to deep space, which acts as a sink at 2.7 K.

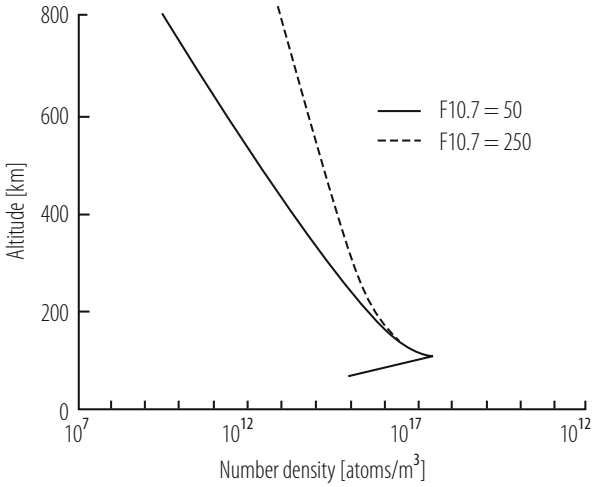


Fig. 18.1 Altitude profiles of number density of atomic oxygen at solar minimum (solid line) and solar maximum (dashed line) [12]

18.2.2.1 Direct Sunlight

A main source of thermal energy is of course the Sun, which acts as a black body at a temperature of 5777 K. The Sun is a very stable source of energy: at the Earth the energy flux varies from 1414 W/m^2 during winter time to 1322 W/m^2 during summer time. The mean intensity at 1 AU is called *solar constant* and is equal to 1367 W/m^2. The spectral energy distribution is approximatively 7% UV, 46% visible and 47% near-IR.

18.2.2.2 Albedo

Albedo refers to the sunlight reflected by a planet. It is highly variable with the conditions of the surface. For spacecraft orbiting close to the Earth, the *albedo* can reach a significant fraction, up to 57%, of the Earth emitted radiation, which in turn is 200–270 W/m^2, depending on the latitude and of the orbit inclination. The Earth itself is a blackbody radiating at around 255 K. This energy cannot be reflected away form the spacecraft which is approximatively at the same temperature. This energy can only be rejected through the spacecraft thermal control system. It is a non negligible amount of radiation: for example, when the Shuttle bay area looks at the illuminated surface of the Earth, its temperature reaches values close to 250 K even if the back of the spacecraft sees the 2.7 K of deep space.

18.2.3 The Charged Component

18.2.3.1 The Low Energy Plasma

At typical Shuttle altitudes, ~300 km, about 1% of the atmosphere is ionized. This fraction increases to 100% at geosynchronous altitudes. This plasma environment can easily charge up satellite components, both on the surface and on the interior of the spacecraft. If the charging exceeds the electric breakdown and discharges are produced they can damage the satellite electronics. The charged component of the radiation is heavily influenced by the existence of the Earth magnetic field. The Earth magnetic field is roughly dipolar:

$$B(R, \theta) = (1 + \sin^2 \theta_M)^{1/2} B_0 / R^3 \tag{18.1}$$

where B is the local magnetic field intensity, θ_M is the magnetic latitude, R is the radial distance measured in Earth radii (R_E) and B_0 is the magnetic field at the equator and at $R = 1$, $B = 0.30$ G. The interaction between the solar wind and the Earth's magnetic field results in a magnetic field structure much more elongated on the night side than it results on the day side, known as *magnetotail*. The resulting magnetic structure is called *magnetosphere* (Fig. 18.2).

The electrical potential of a spacecraft or payload is measured with respect to the nearby plasma when the net charge flow is zero. This current is the sum of the various exchanges of charge between the plasma and the spacecraft including photo-extraction and secondary emission from the spacecraft surfaces. The single

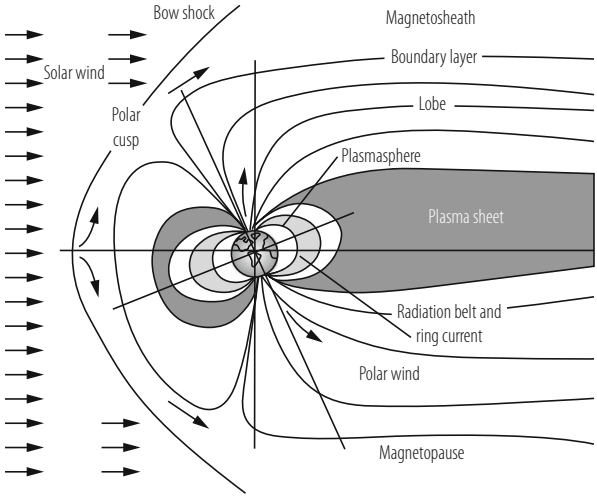


Fig. 18.2 Cross section of the Earth's magnetosphere, showing the key plasma and energetic particle populations [12]

component voltage to the spacecraft ground depends on the element capacitance to the nearby materials. Space charging is particular detrimental in orbits where electron energies in the 10 to 20 keV range dominate the current from the plasma to the spacecraft. At low altitudes this happens only at high latitudes where there are energetic auroral electrons [13]. At other low altitudes locations, low energy electrons are sufficiently abundant to keep the electric fields below the breakdown levels.

The situation is different in higher orbits, such as geosynchronous, where surface charging occurs during magnetospheric substorms between the longitudes corresponding to midnight and dawn [14]. The design of spacecrafts capable to keep a small differential potential with respect to the plasma or to tolerate electrostatic discharges is necessary for these orbits. Design rules and material selection criteria have been developed to help reducing the effect of surface charging on spacecrafts and payloads [15, 16].

It should be noted that, although in the equatorial regions of LEO differential charging is small, the potential of the spacecraft with respect to the surrounding plasma can reach a level close to 90% of the solar array voltage. This should be taken into account when designing experiments aimed to study the plasma properties or when dealing with high voltage power supplies.

18.2.3.2 The Trapped Radiation

Well inside the magnetosphere lie the radiation belts, regions where energetic ions and electrons experience long-term magnetic trapping [17]. Since this trapping requires stable magnetic fields, near the magnetopause the magnetic field fluctuations induced by solar wind prevent long term trapping. On the low altitude side the atmosphere limits the radiation belts to the region above 200 km. The magnetic geometry limits the trapping volume to magnetic latitudes of about 65°. A $magnetic\ L-shell$ is the surface generated by rotating a magnetic field line around the Earth dipole axis and L is measured in units of Earth radius. Trapped particles spiral along paths centered on a given shell. The shell surface can be approximately described as: $R = L\cos^2\theta_M$ [18]. Electrons preferentially populate the toroidal region centered on $L \sim 1.3$ (inner zone) while protons populates the region around $L \sim 5$ (outer zone). The energy of these trapped particles is greater than 30 keV and can reach hundreds of MeV. The intensity of the trapped radiation flux can reach the maximum intensity of $10^8 - 10^9\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ at a distance of $\sim 2\ R_E$ for $E_k > 0.5$ MeV electrons and of $\sim 3\ R_E$ for $E_k > 0.1$ MeV protons. Satellite components, in particular electronics, can be damaged by this penetrating charged form of radiation. A dramatic example of this occurred in 1962 when several satellites ceased to operate after their solar cells were damaged by the increase of radiation belts intensity from high altitude nuclear explosions. Since the basic principles of the trapping are well understood, radiation belts can be modeled quite accurately: a standard model of the Van Allen Belts is available by the National Space Science Data Center [20]. It should be noted, however, that due

to the structure of the Earth magnetic field, which has a dipolar structure not aligned with the Earth angular momentum, the radiation belts are only approximatively of toroidal shape: in the vicinity of the South Atlantic, the structure of the belts is strongly affected and the bouncing altitude of the trapped particles decreases very significantly (South Atlantic Anomaly, SAA). This leads to a region which, although located at LEO altitudes, is characterized by a very intense particle flux, since it is basically within the belts.

Energetic particles, such as electrons from 200 to 1.5 MeV, can implant in the dielectrics and produce discharges within the components themselves (*bulk charging*). At even higher energies, above few MeV, charged particles are highly penetrating and release their energy in the form of ionization deep inside materials. The damages induced by this penetrating radiation can be divided into:

- total dose effects which can degrade the material properties of microelectronics devices, optical elements (lenses, mirrors), solar arrays, sensors, ...
- Single Event Effects or Phenomena (SEE or SEP), effects induced by single particles creating short circuits which can temporarily or permanently damage microelectronics components. They are further subdivided into
  - *Single Event Upset (SEU)* or *bitflip* which although do not damage the electronics may influence the operation of onboard software.
  - *Single Event Latch-up (SEL)*, causing sudden low resistance paths and subsequent drift on the power lines of electronics components which start to operate abnormally until the correct voltage is restored. Depending on the power supply performances SEL can be recovered or could cause permanent damages.
  - *Single Event Burnout* (SEB), causing permanent failures of electronic devices.

18.2.3.3 Solar Particle Events

The Solar Particle Events (SPE) occur in association with solar flares. They consist in an increase of the flux of energetic particles, mostly protons, ($\sim$1 MeV to $\sim$1 GeV) over time scales of minutes, lasting from few hours to several days. Although SPEs occur at a rate of few per year, they are very dangerous for payloads and astronauts, due to the intense radiation dose they deliver, several orders of magnitude higher than in normal conditions (see Fig. 18.3). The global time structure of a SPE is somewhat characteristic (see Fig. 18.4), although the detailed structure depends on the evolution of the original solar flare. X-rays reach the Earth within minutes together with the most relativistic part of the proton spectra; lower energy particles diffuse over time scales of several hours. The fast component of a SPE can be used as early warning to protect the most delicate parts of a payload by switching them off, by using radiation shields or changing the satellite attitude or operational mode.

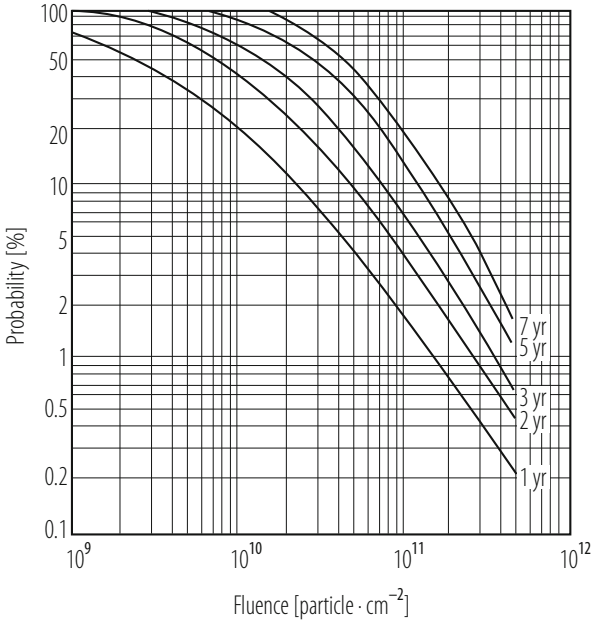


Fig. 18.3 Probability of exceeding a given fluency level as a function of mission duration [12]

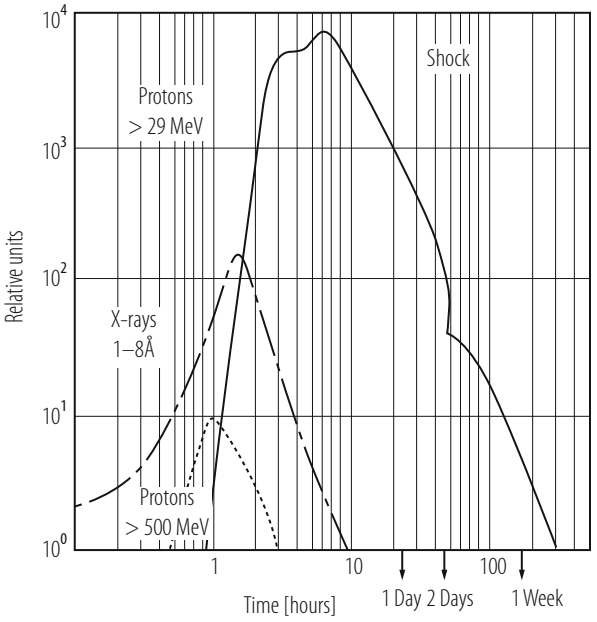


Fig. 18.4 Typical time evolution of a Solar Particle Event (SPE) observed from the Earth [12]

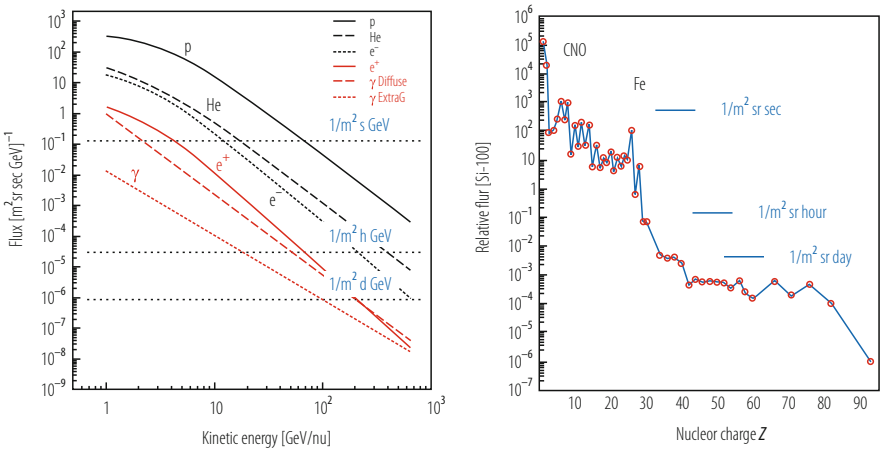


Fig. 18.5 Galactic Cosmic Rays Composition. Left, differential flux of H and He nuclei compared with e^-, e^+ and γ rays. Right, total flux of the nuclear component of galactic Cosmic Rays as a function of the electric charge Z

18.2.3.4 Galactic Cosmic Rays

Galactic Cosmic Rays (GCR) are high energy charged particles reaching the Earth from outside the solar system. The GCR composition is similar to the composition of the energetic particles within the solar system but extend to much higher energy (see Fig. 18.5 and Sect. 18.2.3.4). Their energy ranges from $O(100\,\text{MeV})$ to $10^6\,\text{GeV}$ or more, with an energy spectrum falling as $\sim E^{-2.7}$ for $E > 1\,\text{GeV}$. These particles are very penetrating, loosing their energy only by ionization. Nuclear interaction phenomena are indeed negligible in space for what concerns radiation damages. The ionization losses can create $SEEs$ as discussed above. $GCRs$ have a significant content of high Z particles, fully ionized nuclei with charge extending up to Iron ($Z = 25$). Since ionization losses are proportional to Z^2, high Z GCR can be very effective in causing $SEEs$.

18.2.4 Space Debris

Orbiting spacecraft are subjected to hypervelocity (several km/s) impacts with micron size or larger pieces of dust or debris, both of natural ($micrometeorites$) and artificial ($orbital\ debris$) origin. These impacts can have dramatic effects on a space mission. The probability of a catastrophic impact can be assessed for a given mission and payload. Some measures can be implemented to reduce the effect of the space debris protecting the most important parts with screens made of multilayered materials which can absorb and dissipate the energy of the incoming fragments.

18.3 Types of Orbits

The choice of the orbit heavily influences satellites and payload design.

Many scientific applications are operating on Earth-Referenced Spacecraft orbits. Depending on their typical altitude we talk of *Low Earth Orbits (LEO)*, which are mostly below the Van Allen Belts (typically below 1000 km of height), and of *Geosynchronous Orbits (GEO)* which are well above the Van Allen Belts. Payloads spending substantial time within the Van Allen belts are exposed to high doses of radiation and requires particular care designing and protecting the electronics from SEE and total doses effect.

Table 18.1 shows the types of specialized Earth-Referenced orbits.

Higher orbits are typical of interplanetary missions; for these missions the typical doses received by the satellite payloads are significantly higher than for LEO but lower than within the Van Allen Belts. Far away from the Earth satellites are not anymore shielded from $SPEs$ by the Earth shadow nor by the screening effect of its magnetic field. SEE due to heavy ions and low energy protons should be carefully taken into account when designing the payload electronics.

The space radiation environment remains one of the primary challenges and concerns for space exploration, in particular for deep space missions of long duration, i.e., when the the combined shield due to Earth magnetosphere and atmosphere vanishes. In the inner heliosphere, major sources of radiation are Galactic Cosmic Rays, Solar Particles and Jovian Electrons. Furthermore, in the space nearby Earth particles (mainly electrons and protons) are trapped within the Van Allen radiation belts. Particles populating such a space environment induce single event and cumulative dose in spacecraft materials and, eventually, create electronics hazards.

Table 18.1 Specialized Earth-Referenced orbits

Orbit	Characteristics	Application
Geosynchronous (GEO)	Maintains nearly fixed position over equator	Communication, weather
Sun-synchronous	Orbit rotates so as to maintain approximately constant orientation with respect to Sun	Earth resources, weather
Molniya	Apogee/perigee do not rotate	High latitude communications
Frozen Orbit	Minimizes changes in orbit parameters	Any orbit requiring stable conditions
Repeated Ground Track	Sub orbits repeats	Any orbit where constant viewing angles are desirable

18.4 Space Mission Design

18.4.1 The Qualification Program

Since repairing in space is extremely expensive, if at all possible, designing and building spacecraft and payloads which maximal reliability is a must in the field of space engineering. It follows that quality control is an essential part during the various phases of the program. The $Qualification\ Program$ adds to the cost of the space hardware construction, sometime very significantly, but it makes sure that the program is not headed for failure.

Qualification tests must be designed and implemented to check that the spacecraft/payload can withstand the challenges of launch, deployment and operation in space. Subsystems and components environmental tests include vibration, shock and thermal vacuum, electromagnetic compatibility and radiation hardness.

Although the goal is the same, testing strategies are not unique. There are indeed various testing methods:

- $dedicated\ qualification\ hardware$ (QM): a set of qualification components is built and tested at qualification levels. A set of flight components (FM) is then built and launched after passing a qualification test a lower levels;
- $proto-flight$ approach: a set of flight components is tested at qualification level then assembled into a subsystem or payload which is tested at qualification levels and then launched;
- $similarity$ approach: demonstrate that the components and the environment are identical to previously qualified hardware.

A typical test sequence includes a series of functional tests preceeding/following each environmental test, for example:

- functional test;
- vibration test (levels depending on the mission);
- functional test;
- shock test (levels depending on the mission);
- functional test;
- thermal-vacuum tests, including some functional tests during exposure;
- Electro Magnetic Compatibility (EMC) tests (if required);
- flash X rays with functional tests during exposure (if required);

18.4.2 Vibration and Shock Test

A payload must withstand vibrations caused by the launch vehicle and transmitted through its structural mount. During launch, payload components may experience shocks due to the explosives used for the separation of the various stages. In case reentry is foreseen, they do experience shocks when entering the atmosphere as

well as during the landing phase. In order to understand the dynamical behavior of the payload and of its mounting under these circumstances, Finite Element Analysis (FEA) dynamic and numerical analysis together with Computer Aided Design (CAD) simulation should be performed. In this way it is possible to search for resonances of the mechanical structures, identifying conditions where the material could be stressed or damaged. Following an iterative process the mechanical design of the payload and of its mountings can be improved until all negative margins are eliminated. Dynamic and vibration tests are then performed on a qualification model, using for example an electro-dynamical shaker operating at frequencies between 5 and 3000 Hz, with a spectrum which depends on the mission characteristics. Table 18.2 shows a typical acceleration spectrum expected for payload launched using the Shuttle transportation system. Qualification levels are typically higher by factor 2 to 4. Shock tests are performed using a similar strategy. For example Fig. 18.6 shows shock levels used to simulate the launch of an Alpha-Centaur rocket.

Table 18.2 Maximum expected flight levels for a shuttle mission

	Frequency range	Frequency dependence
X axis	20–58 Hz	0.0025 g^2/Hz
	58–125 Hz	+9 dB/Octave
	125–300 Hz	0.025 g^2/Hz
	300–900 Hz	−9 dB/Octave
	900–2000 Hz	0.001 g^2/Hz
Overall = 3.1 Grms		
Y axis	20–90 Hz	0.008 g^2/Hz
	90–100 Hz	+9 dB/Octave
	100–300 Hz	0.01 g^2/Hz
	300–650 Hz	−9 dB/Octave
	850–2000 Hz	0.001 g^2/Hz
Overall = 3.1 Grms		
Z axis	20–45 Hz	0.009 g^2/Hz
	45–125 Hz	+3 dB/Octave
	125–300 Hz	0.025 g^2/Hz
	300–900 Hz	−9 dB/Octave
	900–2000 Hz	0.001 g^2/Hz
Overall = 3.1 Grms		

18.4.3 Environmental Tests

The environmental qualification campaign of a space component can be divided into three main steps. The first step consists in the development of requirements and constraints related to the payload and to the mission. The second step is

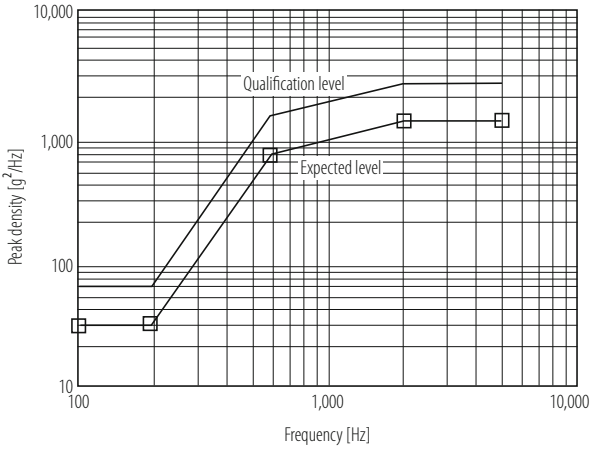


Fig. 18.6 Shock levels simulating the launch environment of an Alpha-Centaur rocket: qualification levels are designed to be greater than the expected design values [12]

to determine and define the space environment (in terms of temperatures, heat transfer ways, worst hot and cold case, etc.) that will characterise thermal conditions throughout the entire life of the component. An important part of the process of qualification is then the thermal analysis which can be conducted using FEA techniques. Once an acceptable thermal model has been developed, test predictions can be calculated to correlate thermal verification tests with the test results. If this correlation is found acceptable, the thermal model is then used to perform flight predictions. If, instead, the correlation is poor the thermal analysis and the hardware configuration need to be carefully checked to understand whether the actual configuration (hardware) requires modifications or the thermal model needs to be updated. Payload temperature requirements derive from the spacecrafts thermal design and the orbital environment and attitude. The purpose of these tests is to demonstrate that the subsystems comply with the specification and perform satisfactorily in the intended thermal environment with sufficient margins. The test environment should be based either on previous flight data, often scaled for differences in mission parameters, or, if more reliable, on analytical prediction or by a combination of analysis and flight data. A margin can include an increase in level or range, an increase in duration or cycles of exposure, as well as any other appropriate increase in severity of the test. Humidity and thermal qualification tests in climatic rooms are performed to test the behaviour of the electronic components and mechanical structures under thermal and humidity changes. The tests are conducted using climatic chambers, with temperature ranges depending on the mission parameters: for a LEO mission typical range lays within $-80\,°\mathrm{C}$ and $+120\,°\mathrm{C}$ for a planetary mission wider intervals are required. Components should be switched on and work both at temperature extremes or during transition, following the mission specifications (see Fig. 18.7).

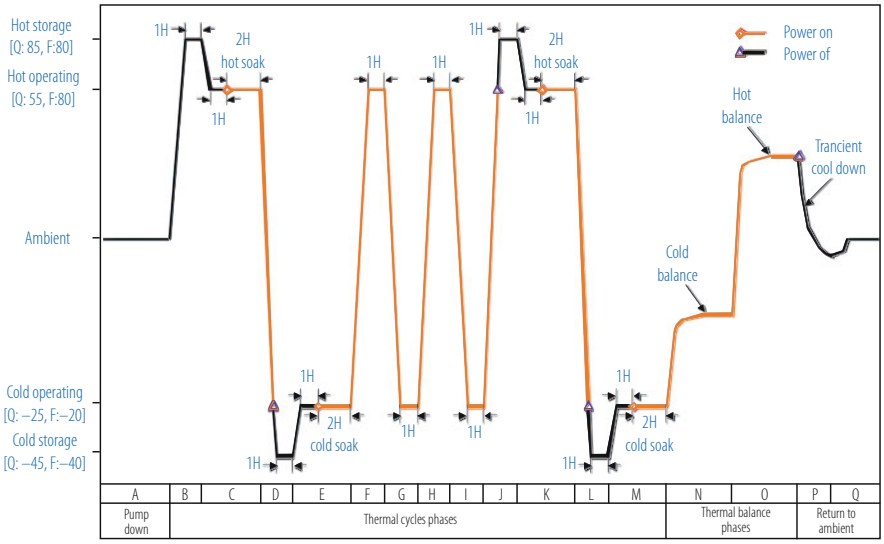


Fig. 18.7 Typical Thermal Vacuum Cycling test profile for a payload to be operated on a LEO orbit on the ISS [54]

18.4.4 EMC Tests

Another area where space payloads are submitted to extensive testing is the compatibility to Electro Magnetic fields, either radiated or received. During the test of radiated EM the device under test is powered and operated in standard operating condition in an EM anechoic chamber. Through suitable antennas and filters read by receivers, the intensity of the emitted radiation is measured as a function of the frequency. The results are compared with the limits requested by specific standards or design rules. If the limits are exceeded, then the electrical grounding or design of the device should be modified. During the received EM test, the device is operated within an EM anechoic chamber while EM radiation, monochromatic or with a specific spectral structure, is generated at a predetermined intensity using special antennas located nearby. The purpose of the test is to check that the item under test does not exhibits anomalies when illuminated by beams of EM radiation, typically emitted by a communication antenna or a nearby electronic device. Figure 18.8 show a typical the result for an EMC radiated test on a payload to be operated on the ISS.

18.4.5 Radiation Hardness Tests

As discussed in the previous paragraphs, the space environment is particularly harsh for operating microelectronics devices, due to the presence of single, heavily

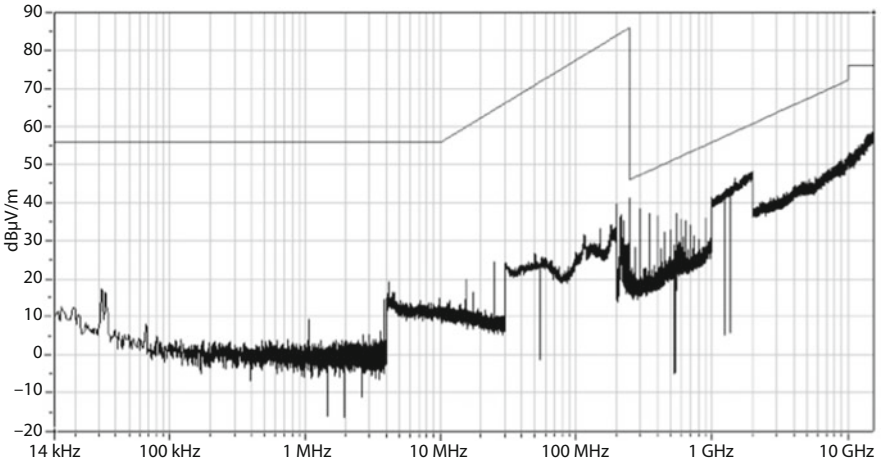


Fig. 18.8 Typical radiated EMC tests result for a payload to be operated on a LEO orbit on the ISS [19]. The thin line presents the e.m. field limits which should not be exceeded as described in Table 18.3

Table 18.3 Radiated EMC limits for the tests described in Fig. 18.8

Frequency range	Emission [dBμV/m]	Antenna
14 kHz–10 MHz	56	rod—vertical
10 MHz–259 MHz	56–86 (16 dB/decade)	biconical—horiz/vert
259 MHz–10 GHz	46–72 (16 dB/decade)	double ridge—horiz/vert
10 GHz–20 GHz	76	horn—horiz/vert

ionizing particles which can deposit large amount of charge in the bulk, inducing short circuits or spurious currents in the solid state circuits. The total dose collected during a space flight is relatively small, mostly in the $krad$ range, so the radiation damages are mostly due to Single Event Effects (SEE).

Depending on the type of circuits and on their construction technology, the sensitivity to ionizing radiation can be very different. In order to select families of commercial circuits which are more insensitive than others it is necessary to run testing campaigns, comparing the behavior of several different chips when exposed to low energy ion beams. The type circuits which shows *latch up* sensitivity or abnormal behavior only at high Linear Energy Transfer (LET) are the one which are radiation hard and can be used in space. In order to ensure statistical significance of the radiation hardness measurement, several chips of each type should be tested (typically > 5).

Often, it is possible to protect the circuit by limiting the current which can flow through the power lines, by the use of an active switch which temporarily cut off the voltage to stop the *latch up* effect. In order to develop and implement a protection scheme it is then important to understand which area of the chip is sensitive. Nowadays it is also possible to perform in laboratory part of these test

by mean of IR laser beams which are absorbed by the silicon and can deposit a controlled amount of energy into the bulk simulating the charge released by a low energy ion [21].

All microelectronics components used in a space experiment must be radiation hard. In addition, the design of the on board electronics should include multiple redundancy since the radiation damage due to SEE is a stochastic process.

Space qualification of radiation resistant devices for a mission not only requires the understanding of damage mechanisms [22], but also the knowledge of local particles (and species) intensities [23, 24] and, in addition, of dose amounts, deposited via ionization and non-ionization energy loss (NIEL) processes. The latter mechanism is that one responsible for displacement damages particularly relevant for semiconductor devices. Only recently, the SR (screened relativistic) NIEL treatment framework has allowed a comprehensive calculation of NIEL doses imparted by electrons, protons, ions and neutrons in any material and compound [25]. SR-NIEL treatment is currently embedded in ESA transport codes, like GRAS [27] and MULASSIS [28] as well as in GEANT4 and it is available at the SR-NIEL and SPENVIS websites.

18.5 Design of a Space Particle Detector

Space born radiation detectors for a space application are, in most cases, adaptation to the space environment of detection techniques used at accelerators or nuclear laboratories.

The environmental conditions discussed in the previous sections obviously influence the detectors design. Particularly important examples are the design of a controlled temperature environment and, for gas detectors, the establishment of controlled pressure conditions.

However, a space born particle experiment has specific limits of different nature which are basically not existing in the case of a laboratory experiment. They are:

- $Weight$. Each kg transported in orbit is very expensive in terms of propellent, costing from 10.000 to 50.000 €/kg, depending on the size of the satellite and orbit of deployment (larger satellites cost less than smaller satellites per kg, higher orbits cost more than lower orbits per kg). This is a substantial limitation for the size of a payload. In addition today space transportation systems have a maximum capacity of about 10 to 20 t in LEO;
- $Power$. The basic source of power in space is the solar energy transformed into electrical power by solar panels. The power can be accumulated in batteries for the periods of the orbit where the spacecraft is shadowed by the Earth. The amount of power consumed by a payload is thus proportional to the area of the panels. One kW of power in space is a large amount of energy consumption. For instance, the entire International Space Station (ISS) power capability does not exceed 110 kW.

- *Volume*. The largest transportation systems can carry payloads which must fit within a cylindrical volume having a maximum radius of about 3 m and a maximum length of about 10 m. Most particle detectors have much smaller sizes. Once in orbit, the size of the payload can increase very significantly, when solar panels, mirrors or radio antennas are expanded from the launch configuration.
- *Accessibility*. Because of the huge cost involved, most of the payload are not accessible during their lifetime in space. Very rare exceptions are the Hubble Space Telescope and the ISS. It follows that the reliability of the instrumentation is essential.
- *Consumables*. Due to the reasons listed above, the amount of consumables is limited. If consumables are needed, e.g. gas for a wire detector or cryogenics for a low temperature payload, the lifetime of the instrument will be limited. In orbit servicing is being developed nowadays for refurbishing the most expensive satellites, but it is still an emerging technology.

These limitations require the ingenuity of the scientist and the knowledge of the engineers to developed most advanced detectors within the available resources.

The reduction of weight calls for the most advanced techniques of CAD (Computer Aided Design) and FEA (Finite Element Analysis) to design structural elements which minimize the amount of material used while tolerating the mechanical stresses and shocks with margins of safety of 2 or more. The techniques used here are typical of aeronautics. The use of light advanced structural materials is mandatory e.g. aluminum, carbon fiber and in general composite materials. Once the structural properties are well defined, static and dynamic FEA is used to identify which part of the structure contribute to the weight without contributing to the structural properties. These parts are normally machined away during the construction. With the advent of Computer Additive Manufacturing (CAM) the weight optimization of structural elements and the integration of functional&structural elements is developing quickly to the advantage of the reduction of the mass of new payloads.

The reduction of power consumption calls for low power electronics and motors. The low power requirement is typical of consumer portable electronics. For this reason modern space experiments make extensive use of electrical devices (VLSI chips, actuators, motors,. . .) used in commercial applications. *Uprating* these parts to be used in space must be a part of the qualification process, in particular from the point of view of radiation hardness, which is not a requirement for consumer electronics. This approach of using $COTS$ (Component Off The Shelf) can reduce significantly the cost of a payload while producing very performant space instrumentation.

Due to the limited accessibility, reliability is a must in space born instrumentation. Reliability is the result of design, manufacturing, integration techniques which must be implemented since the early phases of the development of a payload. During the design phase, redundancy must be implemented in particular in the most critical areas. Special software allows, for example, to measure the probability of the failure of a given circuit, starting from the failure probability of its different components.

Typically the overall probability for a catastrophic failure must be in the range of 1% or less. Single point failures, namely parts of a circuit which are so critical that their failure would generate unacceptable level of malfunctioning, must be avoided. Redundancy of mission critical elements should at least be three to four fold. Similar techniques are applied to test the on board software, exploring all possible software states so to avoid unexpected software conditions which might degrade the payload performances. During the manufacturing phase and integration phases particular care should be given to Quality Assurance (QA), to ensure that the quality of the workmanship of the flight and qualification units and of the fully integrated payload matches the requirements of space standards and specifications. During the testing and qualification campaigns, all possible conditions to be encountered by the payload are simulated to make sure it will operate correctly under any circumstance. QA requires the operators to follow procedures written in advance, perform special tests and report all results and anomalies through written documents which can be verified and used by all the people involved in the various phases of development, commissioning and operation of the payload.

18.6 Space Borne Particle Detectors

The development of modern particle space borne detectors (both for charged particles and photons) has been preceded/accompanied by decades of development of particle detectors for ground based nuclear and particle physics detectors, followed by extended use on stratospheric balloons [29–38].

Small particles detectors have been routinely used on satellites mission to explore the Earth magnetosphere and heliosphere [39, 40].

Modern particle experiments in space can be grouped in three broad categories: (1) experiments measuring the composition, rates and energy spectra of the charged component, (2) experiments detecting single energetic photons and (3) interferometers designed to measure Gravitational Waves in space.

In the first category we find various types of magnetic spectrometers, in the second experiments are based on high granularity tracking calorimeters while the third category cover multiple arms laser interferometers. In the following paragraphs we will briefly discuss some of the most significant space particle detectors developed during the last 10 years, namely AMS-01/02 [41, 42] and PAMELA [43] for the charged component Agile [44] and Fermi [45, 46], for the electromagnetic component and LISA-PF[8] for measuring GW. We will underline the main differences with their ground based counterparts currently used at accelerators experiments. Details of the detection principles, readout electronics or on board software will not be given since they have been addressed in other chapters of this book.

18.6.1 Magnetic Spectrometers

The purpose of a space borne particle detector is to identify the basic properties of the charged cosmic radiation, namely its composition, the energy spectra of the various components and the corresponding fluxes. Thus, the components of a space born magnetic spectrometer are very similar to modern ground based spectrometers, namely:

- a magnet, permanent or superconducting, to measure the sign of the charge by bending the particles path;
- a precise tracking device to measure the particle signed rigidity ($R = B\rho = pc/Ze$), where B is the magnetic field and ρ is the radius of curvature;
- a scintillator based system to trigger the experiment and measure the Time of Flight;
- particle identification (ID) detectors like:
  - Transition Radiation Detectors (TRD) to separate e^+ and e^- from hadrons;
  - Cherenkov Ring Imaging detectors to measure the absolute value of the charge, Z, and the velocity;
  - Electromagnetic Calorimeters to identify the electromagnetic component within the cosmic radiation and measure its energy;
  - Neutron Counters to improve the calorimetric rejection the hadronic CR component.

The first magnetic spectrometers were flown on stratospheric balloons in the 80's. The magnets were based on superconducting coils. The magnets where switched on ground and operated cryogenically for a period of order of 1 day [29, 31–34]. Recently balloon cryostats were able to operate for order of few weeks making possible Long Duration Balloon flights (LDB) around the South Pole [35–38]. Pressurized stratospheric balloons are also beginning to operate Ultra Long Duration Flights (ULDB) which could eventually reach several months duration [90].

The first space borne large magnetic spectrometer, AMS-01 [41] was built only in the mid 90's, due to difficulty of developing a large magnet to be used in space. AMS-01 was the precursor flight of the AMS-02 spectrometer [42], approved by NASA to be flown to and operated on the international space station. (ISS): the engineering model, AMS-01, was operated during the 12 days Shuttle STS91 mission in June 1998 [47]. AMS-02, initially was based on a superconducting magnet to be installed on the ISS in the early 2000's, to be operated for about 3 years, namely for the estimated duration of the superfluid Helium consumable, with the possibility to be reflown after Helium refilling on Earth. The 2003 Challenger disaster forced the earlier retirement of the Shuttle fleet and a modification of the AMS-02 manifest: AMS-02 has been then flown to ISS in 2011 based again on a permanent magnet configuration to benefit of the longest possible exposure ensured by the ISS lifetime. In 2006 a smaller spectrometer, PAMELA [43] also based on a permanent magnet, was launched on a Resurs DK1 Russian satellite to operate in LEO.

One important difference between ground based or balloons magnetic spectrometers and the space borne version is related to the issue of the coupling between the Earth magnetic field and the magnet dipole moment. Since the payload attitude is not a relevant parameter for balloon spectrometers, superconducting magnets exhibiting significant dipole moment can be operated without problems. In space the situation is completely different: the magnetic coupling would affect the attitude of the entire satellite or platform, requiring continuous steering to keep a stable, outward looking attitude. It is then mandatory to design magnets having special geometries (see the following paragraph) and exhibiting minimal magnetic dipole moments.

18.7 Space Spectrometers Based on a Permanent Magnet

All space borne magnetic spectrometers which have been operated in space, AMS-01 (1998), Pamela (2006) and AMS-02 (2011) were based on permanent magnets.

18.7.1 The Alpha Magnetic Spectrometer on Its Precursor Flight (AMS-01)

AMS is an international project involving 16 countries an 56 institutes [42], operated under a NASA-DOE agreement, to install on the ISS a large magnetic spectrometer for the search of nuclear antimatter and to study the origin of dark matter. The first version of the spectrometer was built around a cylindrically shaped, permanent magnet having 800 mm of height and an inner diameter of 1115 mm, resulting in a geometrical acceptance of $0.82\,\mathrm{m}^2$ sr. Figure 18.9 shows the dimensions of the AMS-01 flight magnet. The magnet was made from 64 sectors. Each sector was composed of 100 $5 \times 5 \times 2.5\,\mathrm{cm}^3$ high grade NdFeB blocks. Figure 18.10 shows the arrangement of the field directions of the 64 sectors (left) and the resulting magnetic field map on the middle plane (right). This magnetic configuration is called *magic ring*, and ensures, theoretically, a small magnetic dipole field. To build this magnet the highest grade NdFeB available at the time was used, with an energy level of $(BH)_{max} = 50 \cdot 10^6$ GOe. This configuration resulted in an internal dipole field of 0.15 T and a negligible dipole moment. The total weight of the magnet including the support structure was 2.2 t. The magnetic field, directed orthogonally to the cylinder axis, provided an analyzing power of $BL^2 = 0.15\,\mathrm{Tm}^2$. Outside the magnet the field becomes less then 3–4 G anywhere at a distance larger than 2 m from the magnet center.

Before the construction of full scale magnets, many smaller magnets were built to confirm and measure the field inside the bore, the dipole moment and the flux leakage [41]. Three full scale magnets were built:

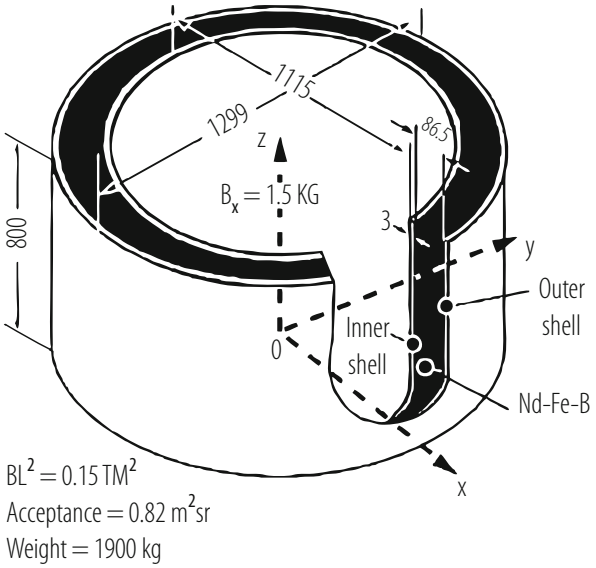


Fig. 18.9 Properties of the AMS-01 flight magnet (dimensions in mm)

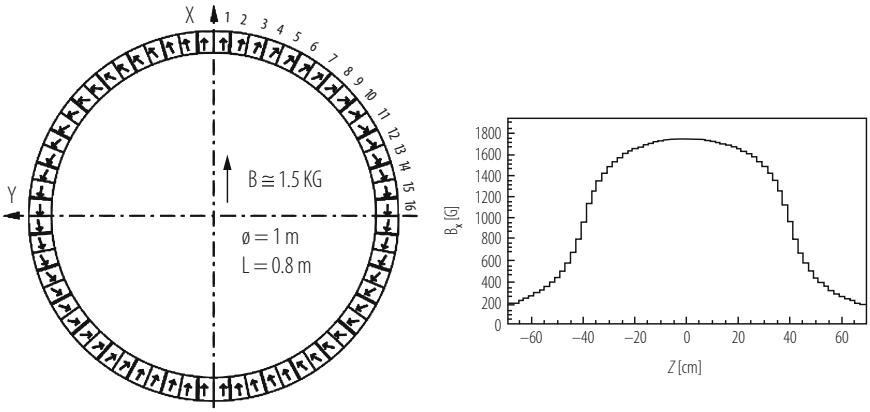


Fig. 18.10 Magnetic field orientation of the AMS-01 magnet sectors (left); B_x field map along the vertical axis ($x = 0$, $y = 0$) (right)

(a) The first magnet was used in acceleration and vibration tests for space qualification.
(b) The second magnet was the flight magnet.
(c) The third magnet was built without glue for NASA safety tests.

The magnet, the supporting structure and space qualification testing were completed by the Institute of Electrical Engineering [48] and the Chinese Academy of Launch Vehicle Technology ($CALT$) [49]. Figure 18.11 shows the first magnet

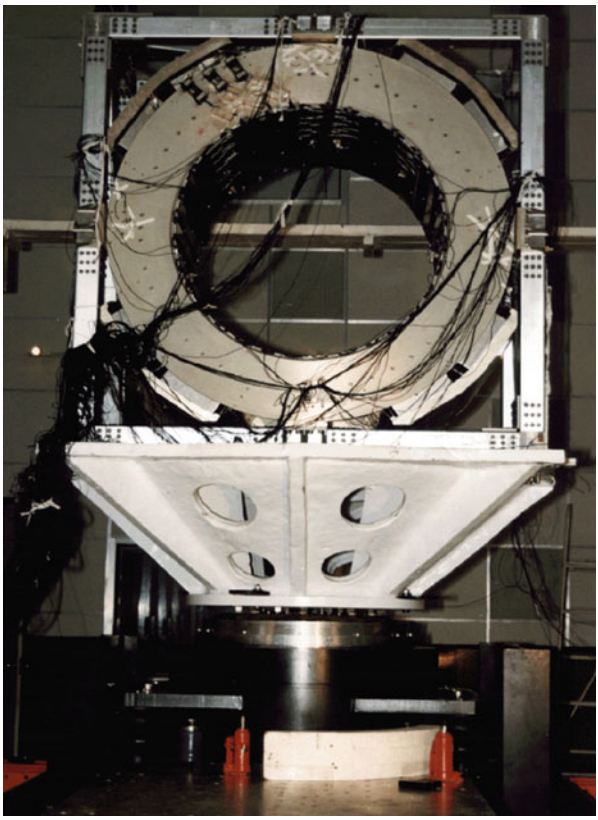


Fig. 18.11 AMS-01 magnet during vibration tests at the Beijing Institute of Spacecraft Environment and Engineering in Beijing, China

Fig. 18.12 AMS-01 magnet undergoing centrifuge (static load) testing at the Laboratory for Centrifugal Modeling in Beijing, China. The picture is blurred since it has been taking through a thick glass window

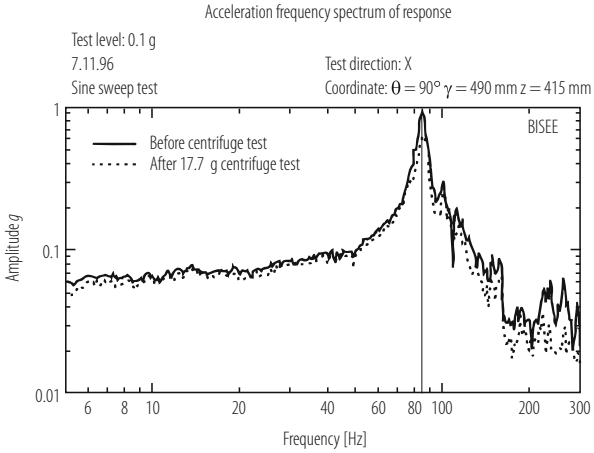


Fig. 18.13 Sine sweep test frequency spectrum response of AMS-01 magnet before and after 17.7 *g* centrifuge test

undergoing vibration testing. Figure 18.12 shows it undergoing centrifuge testing up to 17.7 *g*. Figure 18.13 shows the comparison of the sine sweep test results before and after the 17.7 *g* centrifuge test. The test results indicate that there is no deformation in the detector before and after this test and that the eigenfrequency for the magnet is above the ∼50 Hz region, where the spectral power of the random vibrations produced by shuttle is the highest, as imposed by the NASA safety requirements. The third full scale magnet was built because of the lack of knowledge of the glue performance over an extended period in the space environment. This magnet without any glue was to be tested to destruction to ensure that AMS could be returned on the Shuttle to Earth even if the glue completely failed. The result of the test shows that even with stresses 310 times higher than expected according to analysis the magnet would not break.

During spring of 2006 a smaller but sophisticated magnetic spectrometer, Pamela was launched from Baikonur on a Resource DK Rocket and inserted on a LEO for a 3 years mission. The Pamela experiment is built by an INFN-led international collaboration, and it was launched and operated under an Italian-Russian agreement. The magnet consists of 5 modules of permanent magnets, made of a sintered NdBFe alloy, interleaved by 6 silicon detector planes. The available cavity is 445 mm tall with a section of $1.31 \cdot 10^5\,\mathrm{mm}^2$, giving a geometrical factor of $20.5\,\mathrm{cm}^2$ sr. The mean magnetic field inside the cavity is 0.4 T, providing an analyzing power $BL^2 = 0.1\,\mathrm{Tm}^2$ resulting in a Maximum Detectable Rigidity of 740 GV/c, assuming a spatial resolution of 4 μm along the bending view [43]. The apparatus is 1.3 m high, has a mass of 470 kg and an average power consumption of 355 W. The layout of the magnet and the experiment is shown in Fig. 18.14.

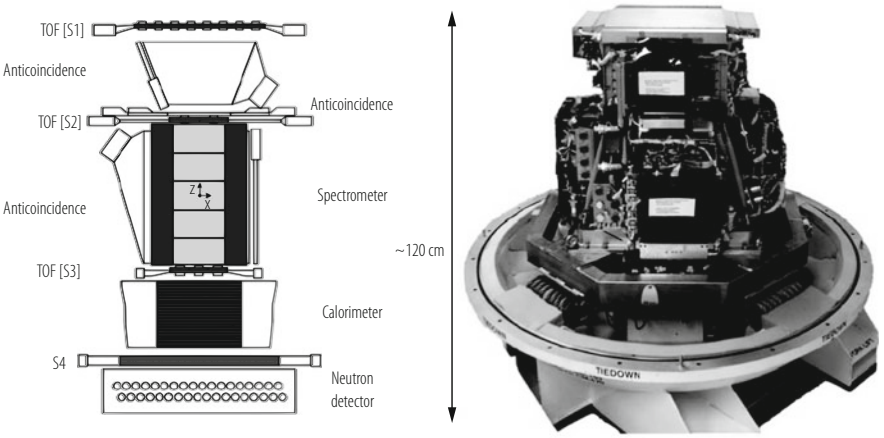


Fig. 18.14 Schematic lateral view of the PAMELA detector (left) and a photograph of it (right) taken before the delivery of the instrument for the integration to the Resours satellite. The geometrical acceptance of the detector is 20.5 cm^2 sr [63]

18.7.1.1 Superconducting Space Spectrometers

The sensitivity to new physics requires spectrometers able to explore higher Cosmic Ray energies while collecting large statistical samples. For this reason ground based modern spectrometers are routinely built using large superconducting magnets which measure particles with momenta in the multi-TeV range [50, 51]. It is of course much more difficult to design a superconducting magnet instead of a permanent magnet to be operated in space. Large facilities like the International Space Station could however provide the necessary infrastructure in terms of power, payload weight and size, data transfer and so on, to install and operate an superconducting spectrometer devoted to high energy particle physics in space. Already in the 80's a proposal was made to install on the Space Station a superconducting spectrometer, ASTROMAG [52]. ASTROMAG was designed around two parallel, large superconducting coils having opposite dipole moments, providing a highly non-uniform magnetic field but an almost zero residual dipole moment. The downsizing of the initial Alpha Station design which took place at the end of the 80's, put the ASTROMAG on indefinite hold status. In 1994 a new proposal was presented through DOE to NASA by the AMS Collaboration, to install and operate a large magnetic spectrometer on the ISS for at least 3 years. This proposal was based on a cylindrical magnetic geometry (*magic ring*), providing much more uniform magnetic field for the particle spectrometer and an almost zero magnetic dipole moment. After the successful flight of the AMS-01 permanent magnet in 1998, the AMS Collaboration proposed to DOE and NASA to upgrade the permanent magnet to a superconducting one having identical geometrical properties but an almost one order of magnitude stronger field (Fig. 18.15).

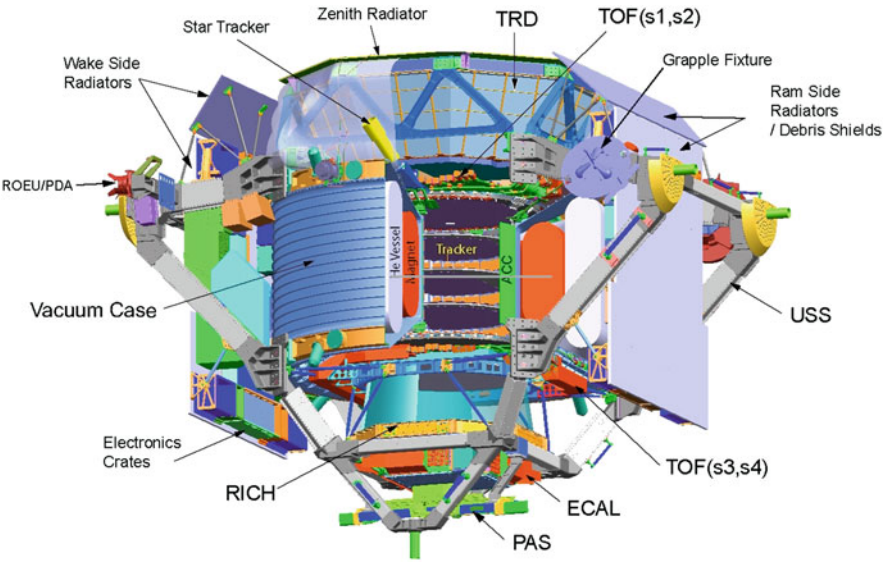


Fig. 18.15 Schematic 3D of the AMS spectrometer in the superconducting version

The project which developed during the years 2000–2010, consisted in the design, construction and extensive testing of the first space qualified superconducting magnet, including thermal-vacuum test in the large ESA-ESTEC space simulator in April 2010. In order to be compatible with the payloads designed for AMS-01, this magnet had identical inner dimensions to the AMS-01 permanent magnet, making the two magnet interchangable with the particle identification detectors. This fact has been instrumental to allow for the switch back to the permanent magnet when it became clear that the early retirement of Shuttle would not have allowed refilling of superfluid ^4He as initially planned. AMS-02 on a permanent magnet configuration has been largely benefitting of the longest possible exposure ensured by the ISS lifetime, which is particularly important in the search of ultrarare events (Fig. 18.16).

The AMS-02 superconducting magnet has been the first designed for operating in space. For this purpose a number of unique challenges had to be solved. Among them:

- endurance: how to maintain the magnet in the superconducting state for the longest possible time, of the order of 3 years, without cryogenic refill;
- safety: how to safely handle the large amount of energy (O(MJ)) stored in the magnet in case of a quench;
- mechanical stability: how to build a structure able to withstand large magnetic forces while being as light as possible.

Two magnets have been built. One is the flight magnet and the other is used for space qualification tests. The magnet system consists of superconducting coils, a

Fig. 18.16 The AMS-02 spectrometer during its integration at CERN in 2009 in the final flight configuration with the permanent magnet

superfluid helium vessel and a cryogenic system, all enclosed in a vacuum tank. The magnet operates at a temperature of 1.8 K, cooled by superfluid helium stored in the vessel. It was designed to be launched at the operating temperature, with the vessel full of 2600 liters of superfluid helium. Four cryocoolers operating between ∼300 and ∼80 K help to minimize the heat losses maximizing the endurance.

The magnet was designed to be launched with no field since it would be charged only after installation on the ISS. Because of parasitic heat loads, the helium will gradually boil away throughout the lifetime of the experiment. After a projected time of 3 years, the helium would be used up and the magnet would warm up and be no longer be operable. Three years of operation in space would indeed correspond to a continuum heat load into the superfluid Helium of about 100 mW, quite a small amount for a magnet which has a volume of about 14 cubic meters.

The coil system consists of a set of 14 superconducting coils arranged, as shown in Fig. 18.17, around the inner cylinder of the vacuum tank. The coil set has been designed to provide the maximum field in the appropriate direction inside the cylindrical bore, while minimizing the stray field outside the magnet. As a result, with the bore geometry identical to the geometry of the AMS-01 magnet, AMS-02 with the superconducting magnet would have had a field almost one order of magnitude larger. A single large pair of coils generates the magnetic dipole field perpendicular to the experiment axis. The twelve smaller flux return coils control the stray field and, with this geometry, they also contribute to the useful dipole field. The magnetic flux density at the geometric centre of the system is 0.73 T. The superconducting wire was developed specifically to meet the requirements of the AMS cryomagnet [53]. The current is carried by tiny (22.4 μm diameter) filaments of niobium titanium (NbTi) which are embedded in a copper matrix, which

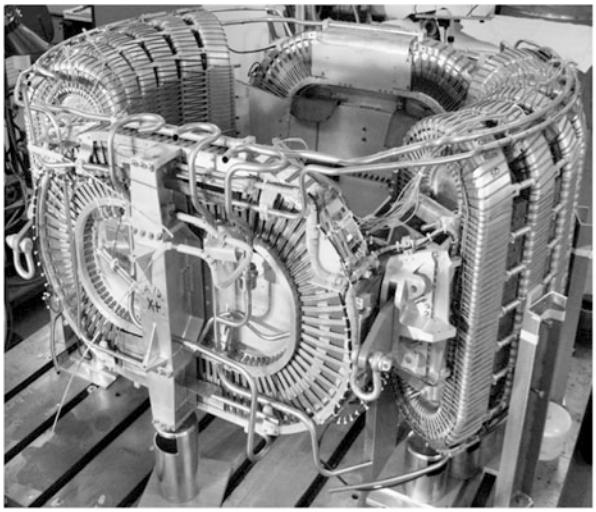


Fig. 18.17 The AMS-02 superconducting magnet: the dipole and the return coils are clearly visible, arranged in the characteristic cylindrical geometry

Table 18.4 AMS-02 superconducting magnet parameters

Parameter	Value
Central magnetic field B_x (at $x = y = z = 0$)	0.750 T
Dipole bending power	0.750 Tm^2
Maximum stray magnetic field at $R = 2.3$ m	13.2 mT
Maximum stray magnetic field at $Y = 2.3$ m	6.62 mT
Maximum stray magnetic field at $R = 3.0$ m	3.4 mT
Peak magnetic field on the dipole coils	5.75 T
Peak magnetic field on the racetrack coils	5.14 T
Maximum torque in geomagnetic field	0.237 Nm
Maximum stray magnetic field at $R = 3.0$ m	3.4 mT
Nominal operating magnet current	400 A
Stored energy	3.72 MJ
Nominal magnet inductance	48 H

is encased in high-purity aluminium. The copper is required for manufacturing reasons, but the aluminium is thermally highly conductive and much less dense, thus providing maximum thermal stability for the same weight. The characteristics of the AMS-02 superconducting magnet are listed in Table 18.4.

The current density in the superconductor is 2300 or 157 $\mathrm{A/mm}^2$ including the aluminium. The 14 coils are connected in series, with a single conductor joint between each pair of adjacent coils. The magnet is designed for a maximum current of 459.5 A, although it is operated at ∼85% of this value. The coils are not coupled thermally. All the coils are constantly monitored by an electronic protection system. If the onset of a quench is detected in any coil, heaters are powered in the other coils

to quench all 14 coils simultaneously. This distributes the stored energy between the coils, preventing any single coil from taking a disproportionate amount of energy which could otherwise result in degradation. The operation of these quench heaters is an important part of the testing and qualification procedure for the magnet coils.

This SC magnet is cooled by superfluid helium, since the thermal conductivity of the superfluid state is almost 6 orders of magnitude higher than in the normal state; in addition, the specific latent heat of the superfluid helium is higher than in normal liquid helium and this can also be used to extend the magnet operation time.

Safety of the AMS magnet had to be assured in ground handling operations, during launch, on orbit and during landing. All cryogenic volumes, as well as the vacuum tank, are protected by burst discs to prevent excessive pressures building up in any fault conditions. Some of the burst discs have to operate at temperatures below 2 K have been the subject of a special development and testing program. In addition, extra protection is provided to mitigate the effect of a catastrophic loss of vacuum. All parts of the AMS magnet system are subject to a battery of tests to ensure their quality, integrity and their suitability for the mission. Every one of the 14 superconducting coils have been tested before assembly into the final magnet configuration. A special test facility has been constructed which allows the coil to be operated under cryogenic conditions as close as possible to the launch. Tests have also been carried out on prototype burst discs. Discs for protecting the vacuum tank have undergone vibration testing followed by controlled bursts. These tests have shown that the discs are not affected by the levels of vibration encountered during a launch. Further tests have been carried out on discs for protecting the helium vessel, which operate at 1.8 K. These discs have been shown to have extremely good leak tightness against superfluid helium.

Mechanical tests of the qualification magnet were done at various facilities: study of the low frequency non-linear behavior were done on a special slip table set up at the SERMS Laboratory [54], in Italy, while static tests were done at IABG [55], in Germany, using a mechanically high fidelity replica of the AMS-02 experiment.

The main characteristics of the AMS01/02 and Pamela magnetic systems are listed in Table 18.5.

Table 18.5 Space borne magnets

Parameter	AMS-01/02	PAMELA	AMS-02[a]
Type of magnet	Permanent	Permanent	Superconducting
MDR [TV]	0.55	0.80	2.6
Magnetic field [T]	0.12	0.48	0.75
Dipole bending power [Tm^2]	0.12	0.085	0.75
Maximum torque in geomagnetic field [Nm]		0.0021	0.24
Maximum geometrical acceptance [cm^2 sr]	5000	20.5	5000

[a]Not deployed in space

Table 18.6 Space borne magnetic spectrometers

Particle ID	AMS-01	PAMELA	AMS-02
Transition radiation detector	No	No	Yes
Time of flight	Yes	Yes	Yes
Silicon tracker	Yes	Yes	Yes
Ring imaging Cherenkov	Yes	No	Yes
Electromagnetic calorimeter	No	Yes	Yes
Neutron counter	No	Yes	No

18.7.2 Particle Identification

High precision study of primary energetic Cosmic Rays requires reliable particle identification. Similar detectors to the one used at the accelerators have been developed and qualified for space usage. With respect to accelerators, however, the task of identification a given particle against its background is significantly different, since, at accelerators, the goal is mostly the identification of short lived particles, while in space short lived particles are irrelevant while the goal is the identification of stable particles and long lived isotopes.

Table 18.6 compare the properties of AMS01/02 and Pamela spectrometers.

18.7.2.1 Tracking Detectors

Silicon detectors, commonly used as tracking devices in ground-based accelerator experiments, offer the best resolution in terms of position measurement. However, a large scale application of these devices in space was never made before AMS-01 [56] in 1998. The AMS-02 silicon tracker [57] (Fig. 18.18) is composed by double-sided micro-strip sensors similar to those used for the L3 [58] micro-vertex detectors at the Large Electron-Positron collider (LEP) at CERN, but the technology and the assembly procedures were qualified for the operation in space. The silicon detectors were produced at Colibrys, SA Switzerland [59] and FBK-irst, Italy [60]. The silicon detectors are assembled together forming ladders up to 60 cm long: particular care was taken to control the readout noise produced by these large silicon assemblies, both from the point of view of the capacitive noise as well as from the point of view of the number of defects, which was requested to be below 10^{-3}. The tracker consists of 8 planes of silicon sensors providing 10 μm (30 μm) position resolution in the bending (non-bending) plane of the 0.15 T field of the magnet. The detectors measure both crossing position and energy loss of charged cosmic ray particles. The readout strips of the silicon sensors are ac-coupled to the low noise, high dynamic range, radiation hard, front-end readout chip, the version Hdr9A of the original Viking design, via 700 pF capacitor chips [61]. Once the charge is known, the momentum is determined by the coordinate measurements in the silicon, which are used to reconstruct the trajectory in the magnet field.

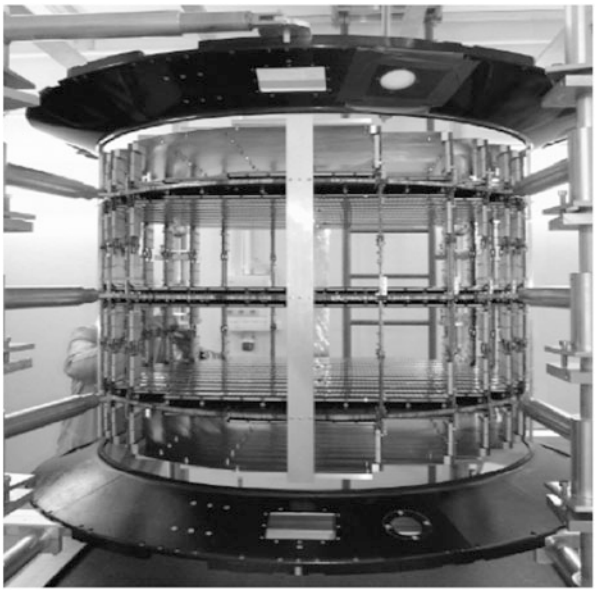


Fig. 18.18 The 8 layers Silicon Tracker of the AMS-02 experiment: the inner planes consists of three double layers of silicon detectors

A similar approach was followed by the PAMELA collaboration. Here the tracking device [62] was based on high accuracy double sided silicon microstrip detectors organized in 12 cm long silicon ladders, produced by Hamamatsu Photonics [64] while low noise, low power, VLSI VA1 chips were used for the front-end section. The use of low-noise front-end electronics is of great importance since the spatial resolution of the detector is strongly related to its signal-to-noise ratio. The applied position finding algorithm gives a spatial resolution of $2.9 \pm 0.1\,\mu\text{m}$ [63]. The junction side shows a larger signal-to-noise ratio ($S/N = 49$) and a better spatial resolution. For this reason this side was used to measure the position along the bending view.

18.7.2.2 Time of Flight Detector

The Time-of-Flight (ToF) measurement is typically associated with the experiment trigger, and, in case of compact magnetic spectrometers, these detectors operates in presence of significant magnetic fields. Figure 18.19 show a schematics of the AMS-02 [65] ToF system, the largest of such systems built to date for space operation. This design follows the experience gained with the AMS-01 detector [66], modified to take into account the different conditions in AMS-02, in particular the stronger stray magnetic field at the photomultiplier tubes (PMTs) which can reach several hundred of G. Each scintillating paddle is instrumented with two PMTs at each

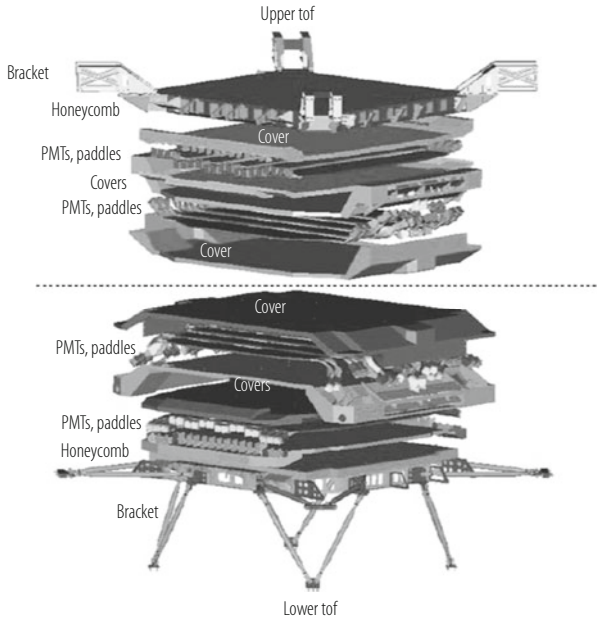


Fig. 18.19 Exploded view of the AMS02 ToF system

end. The time resolution needs to satisfy the physics requirements is 160 ps. The scintillator paddles are 1 cm thick, a compromise between minimum thickness and the light output needed to reach this resolution. Downward going charged particles are distinguished from upward going at the level of 10^9. The system measures the energy loss by a charged particle (to first order proportional to the square of the particle charge) with a resolution sufficient to distinguish nuclei up to charge $Z \sim 20$. Taking into account the attenuation along the counters and the need to have a good measurement of singly charge particles, a dynamic range of more than 10,000 in the measurement of the pulse height is required.

Each paddle is encased in a mechanically robust and light-tight cover and the support structure conforms to the NASA specifications concerning resistance to load and vibrations. The electronics withstands the highly ionizing low Earth orbit environment. Moreover the system guarantees redundancy, with two PMTs on each end of the paddles and double redundant electronics. The system can operate in vacuum over the temperature range -20 to $+50\,^{\circ}$C, it has a weight of less than 280 kg and a power consumption, including all electronics, lower than 170 W. System components have been qualified for use in space and have been extensively tested with particle beams.

18.7.2.3 Transition Radiation Detector

Because of their low mass, Transition Radiation Detectors (TRD) are well suited for utilization in primary Cosmic Ray experiments to separate leptons (electrons) from hadrons (protons) up to hundreds of GeV of energy. The principle of the TRD is very well understood and these detectors are used in large particle physics experiments like ATLAS [67] and ALICE [68] at CERN, and HERA-B at DESY [69]. However, $TRDs$ are gas based detectors and the new challenge is to operate such a large gas detector safely and reliably in space. This has been achieved in the design and construction of the large AMS-02 TRD [70]. The TR photons are detected in straw tubes, filled with a Xe:CO_2 (80%:20%) gas mixture and operated at 1600 V. With a probability of about 50% TR photons are produced in the radiator, 20 mm thick fleece located above each straw layer. Figure 18.20 shows the TRD on top of the magnet vacuum case. The gas tightness of the straw modules is the most critical design issue. The available supplies of gas, 49.5 kg of Xe and 4.5 kg of CO_2, will have to last for 3 years of operation. Using as standard conditions 1 bar and 298 K, this corresponds to 84201 of Xe and 25301 of CO_2. The CO_2 leak rate for one meter of straw-tube was measured to be $0.23 \cdot 10^{-6}$ l mbar/s with the TRD gas Xe:CO_2 80:20 mixture. This leak rate is attributed to diffusion through the straw walls. It corresponds to $1.85 \cdot 10^{-5}$ l mbar/s per module-meter or $9.3 \cdot 10^{-3}$ l mbar/s for the full TRD (500 module meters). A single polycarbonate end piece has a CO_2 leak rate of $0.9 \cdot 10^{-5}$ l mbar/s, for all 328*2 end pieces this totals to $5.9 \cdot 10^{-3}$ l mbar/s. Summing, the total TRD CO_2 leak rate of $1.5 \cdot 10^{-2}$ l mbar/s would correspond to a loss of CO_2 over 3 years of 287 l or a safety factor of 8.8 with respect to the CO_2 supply. This low leak rate has been verified on the completely integrated detector, which could then operate in space for about 26 years. Fabricated TRD modules are accepted if they have a leak rate better than a factor 4 with respect to the overall detector limit. This can only be assured by testing each of the 5248

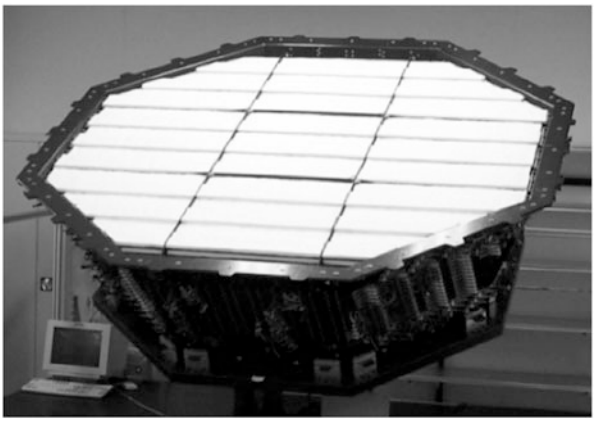


Fig. 18.20 The AMS02 Transition Radiation Detector (TRD) system

straws individually before producing a module [70]. The optimized AMS-02 TRD design with a diameter of 2.2 m and 5248 straw tubes arranged in 20 layers weighs less than 500 kg.

The thermal stability of the TRD is essential for the performance of the detector as temperature variations change the gas density and hence the gas gain. To keep these variations below the 5% level, comparable to other module to module inter-calibration uncertainties, temperature gradients within the TRD should not exceed $\pm 1\,^{\circ}\mathrm{C}$. To keep the spatial and temporal orbit temperature gradient below $1\,^{\circ}\mathrm{C}$ the TRD will be fully covered in multi-layer-insulation (MLI), including the front end electronics. Thermal simulations for orbit parameters which will give the highest TRD temperature swing have been done and prove the effectiveness of this approach. Nonetheless, this has been backed up by a full scale thermal vacuum test in the large volume space simulator at $ESA\ ESTEC$, Holland.

18.7.2.4 Ring Cherenkov Imaging Detector

Cherenkov light is very useful in measuring the velocity and the charge of particles up to tens of GeV of energy, providing a precise measurement to be used together with the momentum determination provided by the spectrometer to identify the different isotopes in the CR flux. The mass of a particle, m, is related to its momentum, p, and velocity, β, through the expression $m = (p/\beta)\sqrt{(1-\beta^2)}$ and its determination is based on the measurement of both quantities. In the AMS spectrometer, the momentum is determined from the information provided by the Silicon Tracker with a relative accuracy of 2% over a wide range of momenta. This entails an error of the same order on the mass of the particle if the velocity is measured with a relative accuracy of about 1 per mil: this is achieved by fitting the shape of the Cherenkov rings measured on the focal plane by high granularity ($4 \times 4\,\mathrm{mm}^2$) pixel photomultipliers located on the focal plane. For this purpose a Ring Imaging Cherenkov Detector (RICH) [71] has been designed with a large geometrical acceptance to operate in the environmental conditions of the outer space. The velocity is determined from the measurement of the opening angle of the Cherenkov cone produced within a radiator layer and the number of detected photons will provide an independent estimation of the charge of the incoming particle.

The measured distribution of charges in the beam is shown in Fig. 18.21 where the structure of individual ion peaks up to $Z = 26$ (Fe) is clearly visible (protons have been suppressed). This spectrum has been fitted to a sum of Gaussian distributions and from their widths we have estimated the charge resolution for each of the ions.

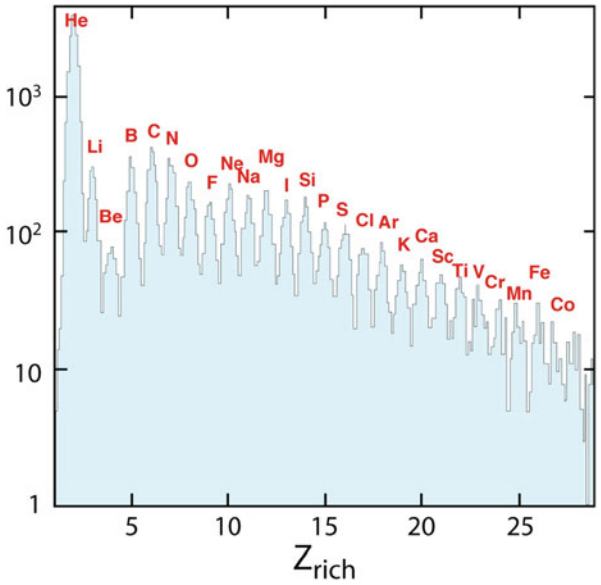


Fig. 18.21 Charge separation of the AMS02 Ring Imaging Cerenkov Detector (RICH) system

18.7.2.5 Electromagnetic Calorimeters

Protons and electrons dominate the positively and negatively charged components of CR, respectively. The main task of the calorimeter is helping the magnetic spectrometer to identify positrons and antiprotons from like-charged backgrounds which are significantly more abundant. Positrons must be identified from a background of protons that increases from about 10^3 times the positron component at 1 GeV/c to $5 \cdot 10^3$ times at 10 GeV/c, and antiprotons from a background of electrons that decreases from $5 \cdot 10^3$ times the antiproton component at 1 GeV/c to less than 10^2 times above 10 GeV/c.

The Electromagnetic Calorimeter ($ECAL$) of the AMS-02 experiment is a fine grained lead-scintillating fiber sampling calorimeter with a thickness corresponding to about 17 radiation lengths [72, 73]. This configuration allows precise, three-dimensional imaging of the longitudinal and lateral shower development, providing at the same time high ($>10^6$) electron/hadron discrimination in combination with the other AMS-02 detectors and good energy resolution, in the range $\sim$1 to $\sim$1000 GeV when the maximum of the e.m. shower is still within the calorimeter. The $ECAL$ also provides a standalone photon trigger capability to AMS. The mechanical assembly has met the challenges of supporting the intrinsically dense calorimeter during launch and landing with minimum weight. The light collection system and electronics are optimized for the calorimeter to measure electromagnetic particles over a wide energy range, from GeV up to TeV.

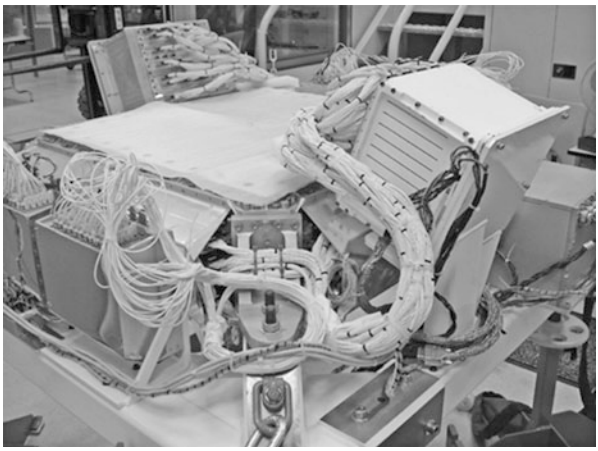


Fig. 18.22 The AMS02 Electromagnetic Calorimeter (ECAL) system

The calorimeter has a total weight of 496 kg. The ECAL mechanical assembly, shown in Fig. 18.22, supports the calorimeter, PMTs and attached electronics. It is designed to minimum weight with a first resonance frequency above 50 Hz, a capability to withstand accelerations up to 14 g in any direction and thermal insulation limiting the gradients (the external temperature ranges from −40 to +50 °C).

The PAMELA ECAL system is a sampling electromagnetic calorimeter comprising 44 single-sided silicon sensor planes (380 μm thick) interleaved with 22 plates of tungsten absorber [74]. Each tungsten layer has a thickness of 0.26 cm, which corresponds to 0.74 X_0 (radiation lengths), giving a total depth of 16.3 X_0 (0.6 nuclear interaction lengths). Each tungsten plate is sandwiched between two printed circuit boards upon which the silicon detectors, front-end electronics and ADCs are mounted. The $(8 \times 8)\,\mathrm{cm}^2$ silicon detectors are segmented into 32 read-out strips with a pitch of 2.4 mm. The silicon detectors are arranged in a 3×3 matrix and each of the 32 strips is bonded to the corresponding strip on the other two detectors in the same row (or column), thereby forming 24 cm long read-out strips. The orientation of the strips of two consecutive layers is orthogonal and therefore provides two-dimensional spatial information ($views$). Figure 18.23 shows the calorimeter prior to integration with the other PAMELA detectors.

More recently other space experiments based on fine grained calorimeters have been developed and are operating in space to study the spectrum of high energy electrons and positrons: CALET [75] on the Japanese segment of the ISS and Dampe [76] on a Chinese satellite.

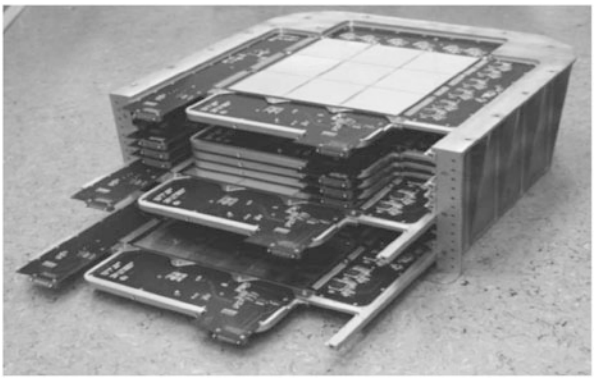


Fig. 18.23 The PAMELA electromagnetic calorimeter. The device is approximately 20 cm tall and the active silicon layer is about $24 \times 24\,\mathrm{cm}^2$. Some of the detecting planes are seen partially, or fully, inserted

18.8 Gamma Rays Detectors

During the last 30 years astrophysicists have discovered the high energy sky, namely sources emitting gamma rays with energy exceeding 1 MeV. The first space borne detector detecting MeV gamma rays were SAS-2 [77] and COS-B [78], followed by the EGRET instrument [79] which extended the energy range to hundreds of MeV with the Compton Gamma-Ray Observatory (CGRO) [80]. More recently Agile [44] and Fermi [45, 46] extended the energy reach to the GeV and hundreds of GeV scale, respectively, closing the gap with the ground based Cherenkov detectors operating from hundreds GeV to tens of TeV.

At these energies the quantized nature of photons is obvious and optical focusing is not anymore possible: high-energy gamma-rays cannot be reflected or refracted and they are detected by their conversion into an e^+e^- pair using techniques developed in nuclear and particle physics. Since both the gamma rays incoming direction and the energy are important informations, the instrument used are a combination of tracking and calorimetric detectors.

EGRET performed the first all-sky survey above 50 MeV and made breakthrough observations of high energy γ-ray blazars, pulsars, delayed emission from Gamma Ray Bursts ($GRBs$), high-energy solar flares, and diffuse radiation from our Galaxy and beyond that have all changed our view of the high-energy Universe. The EGRET instrument (Fig. 18.24), however, was based on detector technologies developed in the 80's: the tracking was provided by a streak chamber while the energy was measured with crystal based NaI calorimeter. In order to eliminate the background due to the charged CRs, about 10^5 times more frequent, the whole instrument was surrounded by a monolithic anti-coincidence counter. This design had two main limitations. First the limited operation time since the tracking device based on a consumable, the gas mixture. Second at increasing photon energy the anti

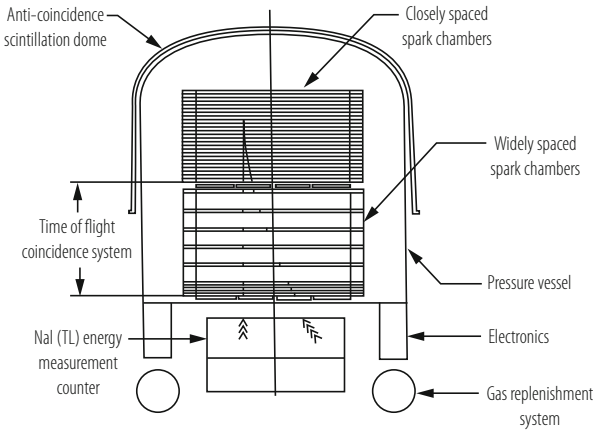


Fig. 18.24 Schematic view of the Energetic Gamma Ray Experiment Telescope (EGRET) on the Compton Gamma Ray Observatory (CGRO)

coincidence system was making the instrument increasingly inefficient due to back scattered particles created in the calorimetric section.

The follow up missions of EGRET, AGILE and Fermi, were based on modern technologies: in these payloads tracking is provided by solid state, imaging calorimeters based on silicon detectors, while the veto system is segmented in several sub elements suitably interconnected within the trigger electronics.

AGILE is a small mission of the Italian Space Agency (ASI), which was launched in April 23rd, 2007. The detector consists on an imaging silicon calorimeter, followed by a thin crystal calorimeter and covered by a coded mask layer to image hard X-rays sources. Its main parameters are listed in Table 18.7.

The Large Area Telescope (LAT) on the Fermi Gamma-ray Space Telescope (Fermi), see Fig. 18.25, formerly the Gamma-ray Large Area Space Telescope (GLAST), was launched by NASA on June 11th, 2008. The LAT is a pair-conversion, high granularity, silicon based imaging telescope made of 16 adjacent towers, followed by an electromagnetic crystal calorimeter. Some of the design choices of Fermi are similar to AGILE, although the detector geometric factor is much larger: each of the 16 Fermi imaging calorimetric towers is equivalent to the whole area of the AGILE detector. In addition the crystal calorimeter section of Fermi is much thicker, providing a much better energy determination. Table 18.8 shows the parameters of the Large Area Telescope instrument.

The self-triggering capability of the LAT tracker is an important new feature of the LAT design made possible by the choice of silicon-strip detectors, which do not require an external trigger, for the active elements [45, 46]. This feature is of essence for the detection of gamma rays in space. In addition, all of the LAT instrument subsystems utilize technologies that do not use consumables such as gas. Upon triggering, the DAQ initiates the read out of these 3 subsystems and utilizes on-board event processing to reduce the rate of events transmitted to ground to a rate

Table 18.7 Agile instrument parameters

Parameter	Value or range
Gamma-ray imaging detector (GRID)	
Energy range	30 MeV–50 GeV
Field of view	~ 2.5 sr
Flux sensitivity ($E > 100$ MeV, 5σ in 10^6 s)	$3 \cdot 10^7$ ph cm^{-2} s^{-1}
Angular resolution	
At 100 MeV (68% cont. radius)	3.5°
At 400 MeV (68% cont. radius)	1.2°
Source location accuracy (high Gal. lat., 90% C.L.)	15 arcmin
Energy resolution (at 400 MeV)	$\Delta E/E \sim 1$
Absolute time resolution	2 μs
Deadtime	∼100–200 μs
Mini-calorimeter	
Energy range	0.35–50 MeV
Energy resolution (at 1.3 MeV)	13% FWHM
Absolute time resolution	∼3 μs
Deadtime (for each of the 30 CsI bars)	∼20 μs

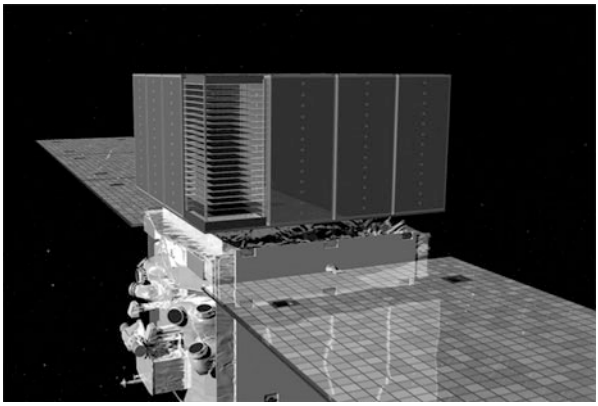


Fig. 18.25 The Fermi Large Area Telescope

compatible with the 1 Mbps average downlink available to the LAT. The on-board processing is optimized for rejecting events triggered by cosmic-ray background particles while maximizing the number of events triggered by gamma-rays, which are transmitted to the ground. Heat produced by the tracker, calorimeter and DAQ electronics is transferred to radiators through heat pipes. The overall aspect ratio of the LAT tracker (height/width) is 0.4, allowing a large field of view and ensuring that nearly all pair conversion events initiated in the tracker will pass into the calorimeter for energy measurement.

Table 18.8 Fermi Large Area Telescope (LAT) parameters [45]

Parameter	Value or range
Energy range	20 MeV–300 GeV
Effective area at normal incidence	9.500 cm^2
Energy resolution (equivalent Gaussian 1σ)	
100 MeV–1 GeV (on axis)	9–15%
1–10 GeV (on axis)	8–9%
10–300 GeV (on-axis)	8.5–18%
>10 GeV (>60° incidence)	≤6%
Single photon angular resolution (space angle)	
On-axis, 68% containment radius	
>10 GeV	≤0.15°
1 GeV	0.6°
100 MeV	3.5°
On-axis, 95% containment radius	$<3 \times \theta_{68\%}$
Off-axis containment radius at 55°	<1.7×(on-axis value)
Field of View (FoV)	2.4 sr
Timing accuracy	<10 μs
Event readout time (dead time)	26.5 μs
GRB location accuracy on-board	$<10'$
GRB notification time to spacecraft	<5 s
Point source location determination	$<0.5'$
Point source sensitivity (>100 MeV)	$3 \cdot 10^{-9}$ ph cm^{-2} s^{-1}

18.9 Gravitational Waves Detectors

Gravitational Waves (GW) are the analogous of the electromagnetic waves for gravitation. They propagate at the speed of light temporarily deforming the texture of space time. Predicted by Albert Einstein [4] on the basis of his theory of General Relativity, gravitational waves transport energy as gravitational radiation, and have been discovered exactly 100 years later by ground based laser interferometers [3]. They are emitted by massive bodies undergoing acceleration. A two body orbiting system, with masses m_1 and m_2, emits a power P:

$$P = \frac{dE}{dt} = -\frac{32}{5}\frac{G^4}{c^5}\frac{(m_1 m_2)^2(m_1 + m_2)}{r^5}. \tag{18.2}$$

Emitted power is really small in most gravitating systems. For example, in the case of the Sun–Earth system, it amounts to about 200 W, about $5 \cdot 10^{-25}$ times less that the electromagnetic power emitted by our star. The GW spectrum extends from frequencies corresponding to the inverse of the age of the universe to few hundreds of Hz (Fig. 18.26).

Their detection has only recently been demonstrated on ground but there are solid reasons to believe that the S/N ratio will be much larger for space borne

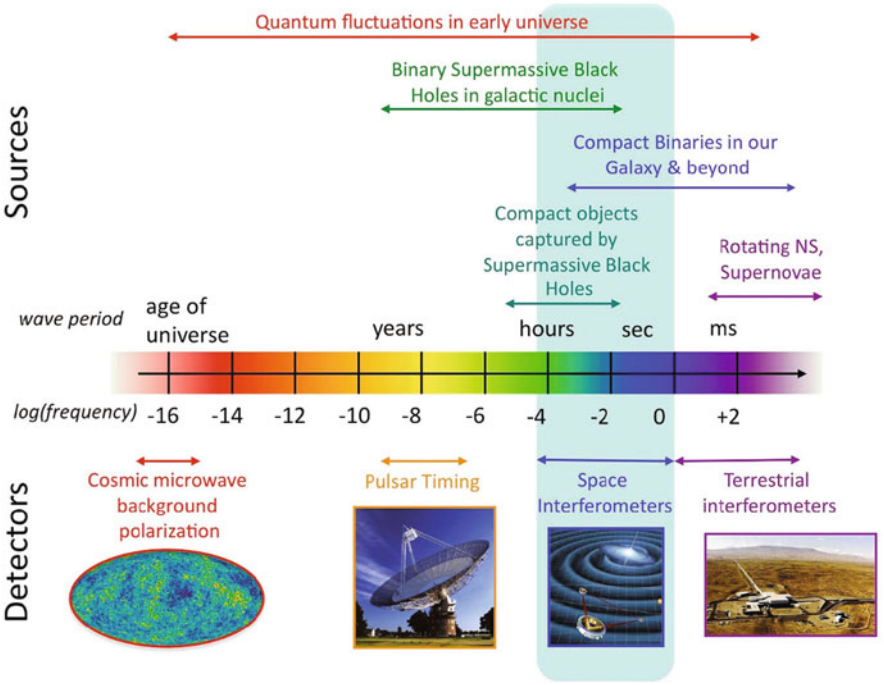


Fig. 18.26 Gravitational wave spectrum and detection techniques

interferometers. LISA is a three-arm space interferometer studied by ESA and NASA up to formulation level for more than 10 years. With the success of the LISA Pathfinder experiment [8], ESA is on track to develop LISA [9] which could be operational towards the beginning of the 30's and detect signals coming from supermassive black-hole mergers, compact objects captured by supermassive black holes and compact binaries (Fig. 18.27).

Once deployed, LISA would measure (a) the orbital period of the binary system, (b) the chirp mass $M = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}$, discriminating between white dwarf, neutron star and black hole binaries and determining the distance for most binary sources with an accuracies better than 1%.

18.9.1 Space-Borne GW Detectors

Measurement of space-time curvature using light beams requires an emitter and a receiver which are perfectly free falling. In flat space-time, the length of proper time between two light-wave crests is the same for the emitter and for the receiver. GW curvature gives oscillating relative *acceleration* to local inertial frames if wave-

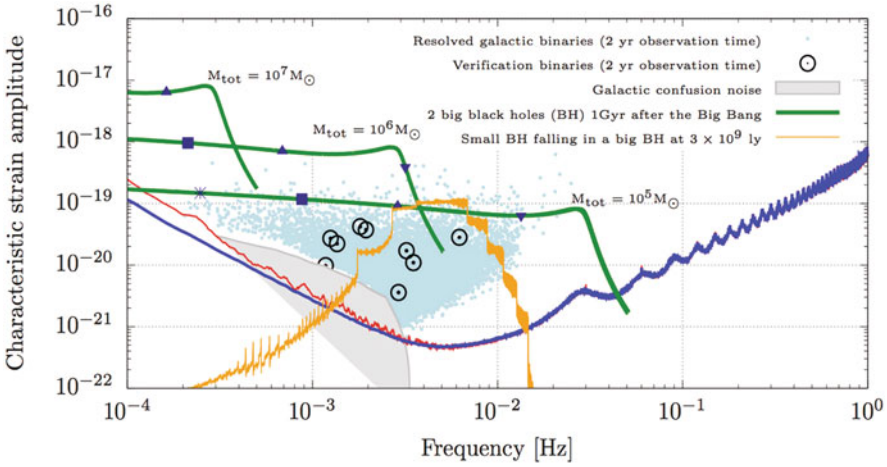


Fig. 18.27 LISA sensitivity to gravitational waves

front is used as a reference: it follows that the receiver sees frequency oscillating. Acceleration of receiver and/or emitter relative to their respective inertial frame produces the same effect of a curvature and should be carefully avoided.

In order to detect gravitational waves via the slowly-oscillating (T up to hours), relative motion they impose onto far apart free bodies, one needs (a) an instrument to detect tiny oscillations, of the size on atom peak-to-peak, ensuring (b) that only gravitational waves can put your test-bodies into oscillation and (c) eliminating all other forces above the weight of a bacteria.

The motion detector (a) is provided by a laser interferometer, as for ground based GW detectors, detecting relative velocities by measuring the Doppler effect through the interferometric pattern variation. Using very stable laser light one can reach the accuracy of 1 atom size in 1 h.

The free falling bodies (b) cannot be touched or supported, at least in an ordinary way. They must be shielded against all other forces (c), in particular, one needs to suppress gravity of the Earth (and of the Sun). The gravity force can be turned off by falling with it, a condition achievable for long periods only on an orbiting satellite. For all other forces, the satellite body would neutralize solar radiation and plasma pressure, actively and precisely following the test mass inertial motion. In order to ensure non contacting (drag-free) behavior, the spacecraft position relative to the test mass is measured by a local interferometer, and it is kept centered on the test mass by acting on micro-Newton thrusters.

The specifications of the LISA GW interferometer design are

- LISA
  - 3 arms, each 5 Mkm
  - $10\,pm/\sqrt{Hz}$ single-link interferometry @1 mHz

– Forces (per unit mass) on test masses $<3\,fm/(s^2\sqrt{Hz})$ @ 0.1 mHz
– 3 non-contacting ("drag-free") satellites

A basic concept of LISA is that the satellites follow independent heliocentric orbits and no formation keeping is needed. In addition the three satellites constellation rotates with respect to the fixed stars providing gravitational waves source location. In the case of the LISA instrument, the implementation of the requirements (a)–(c) is provided by the following main elements:

- the Gravitational Reference Sensor (GRS) with the test mass (also called Inertial Sensor): the GRS is drag-free along sensitive direction, while the other degrees of freedom are controlled via electrostatic forces through a 3–4 mm clearance between test mass and electrodes (Fig. 18.28);
- the Optical Bench with the complete interferometry: it carries all needed interferometry on a monolithic ultra-stable structure obtained by silica hydroxyl bonding (Fig. 18.29);
- a telescope allowing to exchange light with other satellites.

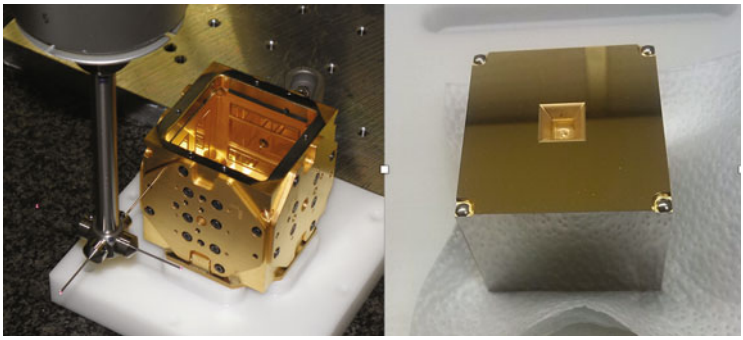


Fig. 18.28 The GRS; left: reference mass housing, right: reference mass

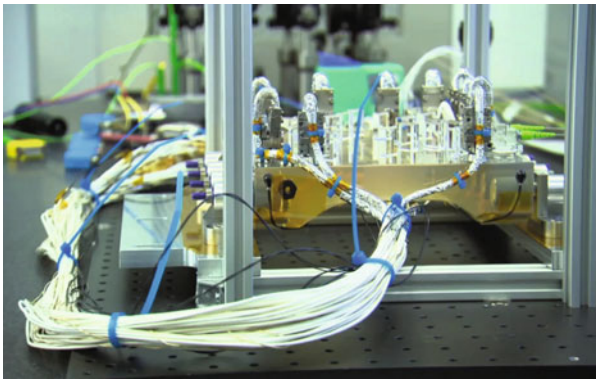


Fig. 18.29 LISA-pathfinder optical bench

18.9.2 LISA Pathfinder

In order to test in space most of the techniques needed for a LISA class space interferometer, the LISA Pathfinder mission has been built, launched in December 3rd 2015 and successfully operated in space during about 8 months, starting from March 1st 2016.

The LISA Pathfinder is based on squeezing of one arm of the final interferometer to within a $O(1)$ m optical bench. This was implemented by removing the long-arm interferometer and replacing the long-arm laser beam reference with a second (quasi-) free test mass. In this miniature implementation of one LISA arm two Au-Pt test masses and two interferometers were placed on the same optical bench. The two masses were not contacting the satellite but the second test mass was forced to follow the first at very low frequency by electrostatic forces (this is different from LISA).

LISA Pathfinder can be seen as a remotely controlled gravitational laboratory operating in space conditions. The GRS consists in two light test masses (2 kg, 46 mm) with a very high density homogeneity ($<<$ 1 μm pores), so that the position of the CoG at geometrical center is known within ± 2 μm. It has a very low magnetic susceptibility $\chi = -(2.3 \pm 0.2) \cdot 10^{-5}$ as well as a negligible magnetic moment $< 4\,\mathrm{nAm}^2$.

Many subtle physical effects apply unwanted forces to test-bodies [81], such as:

- impact with the few molecules that still surround the bodies in high vacuum [82, 83];
- spontaneous electric fields generated by surrounding bodies;
- fluctuating electrical charge from cosmic rays [84];
- changing gravitation generated by thermal deformation of satellite;
- impact with wandering photons;
- fluctuations of the interplanetary magnetic field;

These effects have been studied over the years in the laboratory, pushing forward knowledge in different fields of physics. The results published by the LISA-PF [8] shows that the mission has been very successful, exceeding the predicted accuracy and demonstrating that sub-femto-g differential accelerometry can be achieved, which is an improvement of orders of magnitude with respect to sensors used in the field of experimental gravitation. LISA-PF results confirm the projected LISA sensitivity to the bulk of GW sources present in our galaxy (blue line in Fig. 18.27): a green light for an ESA LISA class mission which could start operating at the beginning of the 30's.

18.10 Future Space Experiments

During the last 20 years an increasing number of modern experiments devoted to particle physics in space have been developed, providing a wealth of new data about CRs composition, high energy astrophysics and gravitational waves. The success of these programs opens the way to the proposal of new, more ambitious projects, designed to measure more accurately the properties of the cosmic radiation.

The universe contains the most powerful particle accelerators, able to accelerate particles to energies inaccessible to ground based laboratories. However these accelerators are quite inefficient and the differential flux of these energetic particles decreases quickly, typically with the third power of the energy. Above a few TeV for the charged component and few hundred GeV for gamma rays, it becomes impractical to develop space instruments having a sufficiently large geometric aperture. For this reason space scientists are considering experiments where the medium where the particle interactions take place is separated from the detector, similarly to what happens for ground based Cherenkov Telescopes, under water or under ice neutrino detectors or Extremely Energetic Cosmic Rays detector arrays, where Cerenkov and fluorescence light produced in the atmosphere, water or ice, respectively, is measured using photon detectors. In the case of these space experiments the medium could be the atmosphere [85, 86], the Moon surface [87] or the magnetosphere [88]: extremely large sensitivities to rare events can be reached by using our whole planet, the Earth, or its satellite, the Moon, as detecting media observable from space borne detectors, collecting emitted light or radio waves by using suitable instrumentation. Discussing these projects is outside the scope of this chapter, however it is interesting to note here a pattern of development which might in the future drive the development of space borne particle experiments devoted to extremely rare events.

18.11 Balloons Experiments

For nearly 40 years, until the mid of the 90's, experiments on stratospheric balloons have been instrumental to study primary CR composition. The advantage of balloons experiments over space experiment is a much lower cost, in the range of 10 M€/mission or less. The main disadvantage is the limited duration of the mission: in the early days it was limited to a day or two, while with the advent of circumpolar flights, the duration has increased to nearly a month/mission. NASA is developing a pressurized balloons technology which would allow for Ultra Long Duration Balloon missions (ULDB) [89, 90] which would reach several months of operation at stratospheric altitudes. In the meantime balloons demonstrated the ability to operate payloads weighting in excess of 1 t, powered by solar panels. It is quite clear that stratospheric balloons missions will be complementary and may become competitive to space missions, in particular when they will last for

several months close to the top of the atmosphere. Most considerations concerning detector developments are quite similar to what has been discussed for space missions: experiments must operate at extreme temperature conditions, withstand shocks, minimize weight and power consumption. Balloons payloads operates in an atmospheric environment, although very rarefied: thermal properties and design should be optimized taking into account also the convective contribution to heat transfer.

References

1. Hess, V., *Über Beobachtungen der durchdringenden Strahlung bei Sieben Freiballonfahrten*, Phys. Z. 13 (1912) 1084.
2. Van Allen, J.A., Ludwig, G.H., Ray, E.C., McIlwain, C.E., *Observations of high intensity radiation by satellites 1958 Alpha and Gamma*, Jet Propulsion 28 (1958) 588–592.
3. Abbott, B. P. et al., *Observation of Gravitational Waves from a Binary Black Hole Merger*, Phys. Rev. Lett. 116 (2016) 061102.
4. Einstein, A.,*Die Feldgleichungen der Gravitation*, Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin, 844–847 (1915).
5. Weber, J., *Gravitational-Wave-Detector Events*, Phys. Rev. Lett. 20 (1968) 1307.
6. Barish, B. C., Weiss, R., *LIGO and the Detection of Gravitational Waves*, Physics Today. 52 (1999) 10.
7. Acernese F. et al., *Advanced Virgo: a second-generation interferometric gravitational wave detector*, Classical and Quantum Gravity, Volume 32, Number 2 (2014).
8. Armano, M. et al., *Sub-Femto-g Free Fall for Space-Based Gravitational Wave Observatories: LISA Pathfinder Results*, P.R.L 116 (2016) 231101.
9. LISA Consortium, *LISA: Laser Interferometer Space Antena*, 20 January 2017.
10. Visentine, J.T. (ed.), *Atomic Oxygen Effects Measurements for Shuttle Missions STS-8 and 41-G*, Vols. I-III, NASA TM-100459 (1988).
11. Leger, L.J., Visentine J.T., Kuminecz, J.F., *Low Earth Orbit Oxygen Effects on Surfaces*, AIAA 22nd Aerospace Sciences Meeting, Reno, NV, Jan. 9–12, 1984.
12. Wertz, J.R., Larson, W.J. (eds.), *Space Mission Analysis and Design*, Microcosm Press and Kluwer Academic Publisher (1999).
13. Gussenhoven, M.S., Hardy, D.A., Rich, F., Burke, W.J., Yeh, H.C., *High Level Spacecraft, Charging in the Low-Altitude Polar Auroral Environment*, J. Geophys. Res. 90 (1985) 11009.
14. Fennel, J.F., Koons, H.C., Leung, M.S., Mizera, P.F., *A Review of SCATHA Satellite Results: Charging and Discharging*, ESA SP-198, Noordwijk, The Netherlands (1983).
15. Purvis, C.K., Garrett, H.B., Witthlesey, A.C., Stevens, N.J., *Design Guidelines for Assessing and Controlling Spacecraft Charging Effects*, NASA Technical Paper 2361 (1984).
16. Vampola, A.L., *The Nature of Bulk Charging and Its Mitigation in Spacecraft Design*, paper presented at WESCON, Anaheim, CA, Oct. 22–24, 1996.
17. Walt, M., *Introduction to Geomagnetically Trapped Radiation*, Cambridge University Press (1994).
18. McIlwain, C.E., *Coordinates for Mapping the Distribution of Magnetically Trapped Particles*, J. Geophys. Res. 66 (1961) 3681–3691.
19. Cervelli, F., et al., *The space qualified read-out electronics for the e.m. calorimeter (ECAL) of the AMS-02 experiment*, IEEE, TNS-00184-2009.
20. National Space Science Data Center web site: http://nssdc.gsfc.nasa.gov/.
21. Alpat, B., et al., *A pulsed nanosecond IR laser diode system to automatically test the Single Event Effects in Laboratory*, Nucl. Instrum. Meth. A 485 (2002) 183–187.

22. see Chapter 3 on *Managing Space Radiation Risk in the New Era of Space Exploration*, Committee on the Evaluation of Radiation Shielding for Space Exploration of the Aeronautics and Space Engineering Board (National Research Council, USA), National Academies Press, Washington DC (2008), ISBN 9780309113830.
23. SPENVIS, *ESA's SPace ENVironment Information System* (2018), available at https://www.spenvis.oma.be/.
24. OMERE software (2018), *Outil de Modèlisation de l'Environnement Radiatif Externe*, the code is developed by TRAD with the support of the CNES and is available at http://www.trad.fr/en/space/omere-sotftware/.
25. SR-NIEL Calculator: *Screened Relativistic (SR) Treatment for Calculating the Displacement Damage and Nuclear Stopping Powers for Electrons, Protons, Light- and Heavy- Ions in Materials* by Boschini, M.J., Rancoita, P.-G. and Tacconi, M., current version 3.9.3 (October 2017) is available at http://www.sr-niel.org/; the treatment can be comprehensively found in Chapters 2, 7 and 11 of [26].
26. Leroy, C. and Rancoita, P.-G., *Principles of Radiation Interaction in Matter and Detection* 4th Edition, World Scientific (Singapore) 2016, ISBN 9789814603188.
27. GRAS *(Geant4 Radiation Analysis for Space)* code is available at ESA website upon registration; the original article is by Santin, G., Ivanchenko, V., Evans, H., Nieminen, P. and Daly, E., IEEE Trans. Nucl. Sci. 52, Issue 6, 2005, pp 2294–2299.
28. MULASSIS - *MUlti-LAyered Shielding SImulation Software, available at ESA website*: the original article is by Lei, F., Truscott, R.R., Dyer, C.S., Quaghebeur, B., Heynderickx, D., Nieminen, P., Evans, H. and Daly, E., IEEE Transactions on Nuclear Science Vol 49 No 6 (2002) P2788–2793.
29. Streitmatter, R.E., *ISOMAX: A Balloon-borne Instrument to Study Berlyllium and Other Light Isotopes in the Cosmic Radiation*, Proc. 23th Int. Cosmic Ray Conf., Calgary 1993.
30. Mitchell, J.W., et al., *(IMAX) Isotope Matter-Antimatter Experiment*, Proc. 23rd Int. Cosmic Ray Conf., Calgary 1993, Vol. 1, p. 519.
31. Carlson, P., Francke, T., Suffert, M., Weber, N., *A RICH counter for antimatter and isotope identification in the cosmic radiation*, Proc. 23th Int. Cosmic Ray Conf., Calgary 1993, Vol. 2, p. 504.
32. Yamamoto, A., et al., *Balloon-Borne Experiment with a Superconducting Solenoidal Magnet Spectrometer*, Adv. Space Res. 14(2) (1994) 75–87.
33. Barwick, S.W., et al., *The High-Energy Antimatter Telescope (HEAT): an instrument for the study of cosmic-ray positrons*, Nucl. Instrum. Meth. A 400 (1997) 34–52.
34. Beatty, J.J., et al., *Cosmic Ray Energetics And Mass (CREAM): A Detector for Cosmic Rays near the Knee*, Proc. 26th Int. Cosmic Ray Conf., Salt Lake City 1999, Vol. 5, pp. 61–64.
35. Isbert, J., et al., *ATIC, a Balloon Borne Calorimeter for Cosmic Ray Measurements*, Proc. 10th Int. Conf. Calorimetry in Particle Physics, Pasadena, CA, March 25, 2002, pp. 89–94.
36. Boyle, P., et al., *Cosmic Ray Energy Spectra of Primary Nuclei from Oxygen to Iron: Results from the TRACER 2003 LDB Flight*, 30th Int. Cosmic Ray Conf., Merida, Mexico (2007).
37. Yoshimura, K., et al., *The First BESS-Polar Flight over Antarctica*, Proc. 25th Int. Symp. Space Technology and Science, Kanazawa, Japan (2006), pp. 1132–1137.
38. Seo, E.S., et al., *CREAM: 70 days of flight from 2 launches in Antarctica*, Advances in Space Research 42 (2008) 1656–1663.
39. Baker, D.N., Mason, G.M., Figueroa, O., Colon, G., Watzin, J.G., Aleman, R.M., *An Overview of the Solar, Anomalous, and Magnetospheric Particle Explorer (SAMPEX) Mission*, IEEE Trans. Geosci. Remote Sens. 31 (1993) 531–541.
40. ESA's Report to the 30th COSPAR Meeting, Hamburg, Germany, July 1994, European Space Agency, Paris, (1992) 47–57.
41. Ahlen, S.P., et al., *An Antimatter spectrometer in space*, Nucl. Instrum. Meth. A 350 (1994) 351–367.
42. AMS Collaboration, Aguilar, M., et al., *The Anti Matter Spectrometer (AMS-02): A particle physics detector in space*, Nucl. Phys. Proc. Suppl. 166 (2007) 19–29.

43. Bonvicini, V., et al., *The PAMELA experiment in space*, Nucl. Instrum. Meth. A 461 (2001) 262–268.
44. Tavani, M., et al., Astron. Astrophys. 502 (2009) 995.
45. Atwood, W.B., et al., *The Large Area Telescope on the Fermi Gamma-ray Space Telescope Mission*, Astrophys. J. 697 (2009) 1071–1102.
46. Meegan, C., et al., *The Fermi Gamma-Ray Burst Monitor*, Astrophys. J. 702 (2009) 791–804.
47. AMS Collaboration, Aguilar, M., et al., *The Alpha Magnetic Spectrometer (AMS) on the International Space Station. Part I: Results from the Testflight on the Space Shuttle*, Physics Reports 366 (2002) 331–405.
48. Institute of Electrical Engineering, IEE, Chinese Academy of Sciences, 100080 Beijing, China.
49. Chinese Academy of Launching Vehicle Technology, CALT, 100076 Beijing, China.
50. CMS Physics, Technical Design Report, Volume I: CERN-LHCC-2006-001, Feb. 2, 2006.
51. ATLAS detector and physics performance, Technical Design Report, Volume I, May 25, 1999.
52. Jones, W. V., *Astromag - Particle astrophysics magnet facility for Space Station Freedom*, IAF, 40th Int. Astronautical Congress, Malaga, Spain, Oct. 7–13, 1989.
53. Blau, B., et al., Grav. Cosmol. Suppl. 5 (2000) 1; IEEE Trans. Appl. Supercond. 12 (2002) 349.
54. SERMS, Via Pentima 4, 05100 Terni, Italy; Bertucci, B., *The S.E.R.M.S. laboratory. A research and test facility for space payloads and instrumentation*, Memorie della Societa Astronomica Italiana 79 (2008) 818.
55. IABG mbH, Einsteinstrasse 20, 85521 Ottobrunn, Germany.
56. Battiston, R., *A silicon tracker for the antimatter spectrometer on the International Space Station ALPHA*, Proc. 1st Arctic Workshop Future Physics and Accelerators, Saariselka, Finland, Aug. 21–26, 1994, (1994) 138–156; Alcaraz, J., et al., *A silicon microstrip tracker in space: Experience with the AMS silicon tracker on STS-91*, Nuovo Cimento A 112 (1999).
57. Alcaraz, J., et al., *The alpha magnetic spectrometer silicon tracker: Performance results with protons and helium nuclei*, Nucl. Instrum. Meth. A 593 (2008) 376–398, Erratum: *ibid.* 597 (2008) 270.
58. Acciarri, M., et al., *The L3 silicon microvertex detector*, Nucl. Instrum. Meth. A 351 (1994) 300–312.
59. Colibrys (Switzerland) Ltd, Maladière 83, 2000 Neuchâtel, Switzerland.
60. FBK-irst, Via Sommarive, 18, 38050 Povo (Trento), Italy.
61. Toker, O., et al., Nucl. Instrum. Meth. A 340 (1994) 572.
62. Picozza, A., et al., Astroparticle Phys. 27 (2007) 296–315.
63. Straulino, S., et al., *Spatial resolution of double-sided silicon microstrip detectors for the PAMELA apparatus*, Nucl. Instrum. Meth. A 556 (2006) 100–114.
64. 5000, Hirakuchi, Hamakita-ku, Hamamatsu City, Shizuoka Pref., 434-8601, Japan.
65. Bindi, V., et al., *The AMS-02 time of flight system. Final design*, Proc. 28th Int. Cosmic Ray Conf., Tsukuba, Japan, July 31 - Aug. 7, 2003.
66. Baldini, L., *The AMS time-of-flight system*, Proc. 27th Int. Cosmic Ray Conf., Hamburg, Germany, Aug. 7–15, 2001.
67. The ATLAS TRT collaboration, Abat, E., et al., J. Instrum. 3 (2008) P02014.
68. ALICE TRD Collaboration, *The ALICE transition radiation detector*, Nucl. Instrum. Meth. A 502 (2003) 127–132.
69. Saveliev, V., *The HERA-B Transition Radiation Detector*, Nucl. Instrum. Meth. A 408 (1998) 289–295.
70. Siedenburg, T., et al., *A transition radiation detector for AMS*, Nucl. Phys. Proc. Suppl. 113 (2002) 154–158.
71. Casaus, J., et al., *The AMS RICH detector*, Nucl. Phys. Proc. Suppl. 113 (2002) 147–153.
72. Adinolfi, M., et al., *The KLOE electromagnetic calorimeter*, Nucl. Instrum. Meth. A 482 (2002) 364–386.
73. Cadoux, F., et al., *The AMS-02 electromagnetic calorimeter*, Nucl. Phys. Proc. Suppl. 113 (2002) 159–165.
74. Bonvicini, V., et al., *A silicon-tungsten imaging calorimeter for PAMELA*, Proc. 26th Int. Cosmic Ray Conf. (ICRC 99), Salt Lake City, Utah, Aug. 17–25, 1999, Vol. 5, pp. 187–190.

75. Torii S. et al., *Calorimetric electron telescope mission. Search for dark matter and nearby sources*, Nucl. Instr. and Meth. A 630 (2011) 55-7; Torii S. et al., Proc. of 33rd ICRC (2013) 245
76. Chang J et al., *Dark Matter Particle Explorer: The First Chinese Cosmic Ray and Hard γ-ray Detector in Space*, Chin. J. Space Sci. 34 (2014); Chang J et al., *The DArk Matter Particle Explorer mission*, Astropart. Phys. 95 (2017) 6
77. Derdeyn, S.M., Ehrmann, L.H., Fichtel, G.J., Kniffen, D.A., Ross, R.W., Nucl. Instrum. Meth. 98 (1972) 557.
78. Bignami, G.F., et al., Space Sci. Instrum. 1 (1975) 245.
79. Thompson, D.J., et al., Astrophys. J. 415 (1993) L13.
80. Thompson, D.J., et al., Astrophys. J. Suppl. Ser. 86 (1993) 629.
81. Carbone, L. et al., *Thermal gradient-induced forces on geodesic reference masses for LISA*, P.R.L. D76 (2007) 102003
82. Carbone L., et al., *Achieving Geodetic Motion for LISA Test Masses: Ground Testing Results* P.R.L. 91 (2003) 151101; Erratum-ibid. P.R.L. 91 (2003) 179903
83. Cavalleri, A., *Increased Brownian Force Noise from Molecular Impacts in a Constrained Volume*, P.R.L. D103 (2009) 140601
84. Antonucci, E. et al., *Interaction between Stray Electrostatic Fields and a Charged Free-Falling Test Mass*, P.R.L. D108 (2012) 181101
85. G. D'Ali Staiti, G., et al., *EUSO: A space mission searching for extreme energy cosmic rays and neutrinos*, Nucl. Phys. Proc. Suppl. 136 (2004) 415–432.
86. Takahashi, Y., *A Giant natural TPC (500 km)^3 to observe extremely high energy cosmic particles - JEM EUSO telescope on International Space Station*, J. Phys. Conf. Ser. 65 (2007) 012022.
87. Battiston, R., Brunetti, M.T., Cervelli, F., Fidani, C., Menichelli, M., *A Moon-borne electromagnetic calorimeter*, Astrophys. Space Sci. 323(4) (2009) 357–366.
88. Gusev, A.A., et al., *Detector for electron spectrum measurements in TeV region on synchrotron radiation in geomagnetic field*, Proc. 21st Int. Cosmic Ray Conf., Adelaide 1990, Vol. 3, pp. 245–248; Anderhub, H., et al., *Preliminary results from the prototype Synchrotron Radiation Detector on Space Shuttle mission STS-108*, Nucl. Phys. Proc. Suppl. 113 (2002) 166–169.
89. *NASA Stratospheric Balloons Pioneers of Space Exploration and Research*, Report of the Scientific Ballooning Planning Team, Oct. 2005; http://sites.wff.nasa.gov/code820/uldb.html.
90. Wiencke, L., *EUSO-Balloon mission to record extensive air showers from near space*, PoS ICRC 2015, (2016) 631.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 19
Cryogenic Detectors

Klaus Pretzl

19.1 Introduction

Most calorimeters used in high energy physics measure the energy loss of a particle in form of ionization (free charges) or scintillation light. However, a large fraction of the deposited energy in form of heat remains undetected. The energy resolution of these devices is therefore mainly driven by the statistical fluctuations of the number of charge carriers or photoelectrons involved in an event. In contrast, cryogenic calorimeters are able to measure the total deposited energy including the heat in form of phonons or quasi-particles in a superconductor. With the appropriate phonon or quasi-particle detection system much higher energy resolutions can be obtained due to the very large number of low energy quanta (meV) involved in the process. This feature makes cryogenic calorimeters very effective in the detection of very small energy deposits (eV) with resolutions more than an order of magnitude better than for example semiconductor devices.

During the last two decades cryogenic detectors have been developed to explore new frontiers in physics and astrophysics. Among these are the quest for the dark matter in the universe, the neutrinoless double beta decay and the mass of the neutrino. But other fields of research have also benefited from these developments, such as astrophysics, material and life sciences.

The calorimetric measurement of deposited energy in an absorber dates back to 1878, when the American astronomer S.P. Langley invented the bolometer [1]. With this device he was able to measure the energy flow of the sun in the far infrared region of the spectrum and to determine the solar constant. Since then the bolometer has played an important role to measure the energy of electromagnetic radiation

K. Pretzl (✉)
Laboratory for High Energy Physics - Albert Einstein Center for Fundamental Physics, University of Bern, Bern, Switzerland
e-mail: pretzl@lhep.unibe.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_19

of celestial objects. At the turn of the century radioactivity was discovered and P. Curie and A. Laborde made a first attempt in 1903 to measure the energy released in radioactive decays using a calorimetric device [2]. Thereafter micro-calorimeters were developed by C.D. Ellis and A. Wooster in 1927 [3] and independently by W. Orthmann and L. Meitner in 1930 [4] to determine the average energy of the electron in the beta-decay of ^{210}Bi. The differential micro-calorimeter developed by W. Orthmann allowed to measure heat transfers of the order of μW. Using this true calorimetric technique, he and L. Meitner were able to determine the average energy of the continuous beta spectrum in ^{210}Bi to 0.337 MeV with a 6% accuracy. These measurements contributed greatly to the notion of a continuous beta-spectrum leading to W. Pauli's neutrino hypothesis in 1930.

In 1935 F. Simon [5] suggested to measure the energy deposited in radioactive decays with low temperature calorimeters. He claimed that with a calorimeter of 1 cm^3 tungsten in a liquid helium bath at 1.3 K, one could measure a heat transfer of nW, which is about 1000 times more sensitive than the calorimeter of W. Orthmann. The argument is that at low temperatures the heat capacity C of a micro-calorimeter is low and a small energy loss E of a particle in the calorimeter can lead to an appreciable temperature increase $\Delta T = E/C$. Later in 1949, D.H. Andrews, R.D. Fowler and M.C. Williams [6] reported the detection of α-particles from a Po source with a bolometer made of a superconducting strip of NbN mounted on a copper base. The operating temperature was chosen 15.5 K, which corresponded to the center of the transition halfway between the superconducting and the normal state of NbN. However, at this stage of the experiment no energy information of the alpha particles could be extracted from the signals, since the signal to background ratio was not sufficient. Their bolometer was used only as a particle counter. In 1969, G.H. Wood and B.L. White [7] were able to measure the energy of the emitted alpha particles from a polonium source with a superconducting tunnel junction (STJ). The energy was derived from the tunneling current, which is proportional to the excitations of the quasi-particles induced by the energy loss of the α-particles in the junction.

H. Bernas et al. [8] introduced 1967 superheated superconducting granules (SSG) to measure beta radiation. Used as an energy threshold detector the energy loss of an electron in a granule could suffice to drive the granule from a super-conducting into a normal state. This phase transition would induce a signal in a pickup coil due to the Meissner effect, provided the granules are sitting in an external magnetic field. A. Drukier and C. Vallette [9] were able to detect charged particles with a SSG device. Later in 1984, A. Drukier and L. Stodolsky [10] suggested the use of SSG detectors for neutrino and astrophysics experiments.

In early 1970 a new type of bolometer, the so-called composite one, was developed by N. Coron, G. Dambier and J. Leblanc [11]. It consisted of an absorber and a thermally coupled thermometer in form of a semiconductor thermistor. Later in 1974, T. Niinikosky and F. Udo [12] proposed cryogenic calorimeters for the detection of neutrinos. E. Fiorini and T. Niinikoski [13] explored in 1984 the possibility of using low temperature bolometers to improve the limits on neutrinoless double beta decays. At this time D. McCammon, S.H. Moseley, J.C. Mather and R. Mushotzky [14] published first results with a cryogenic calorimeter

for X-ray spectroscopy. In 1985 N. Coron et al. [15] developed a cryogenic composite bolometer as a charged particle spectrometer.

Based on all these interesting ideas and developments, a first workshop on low temperature detectors (LTD1) was held in 1987 at Ringberg-Castle on Lake Tegernsee in southern Bavaria. Due to the success of this workshop and the growing interest in this field, further workshops have been organized in Europe, the USA and Japan. Much of the original work in this field can be found in the proceedings of these LTD workshops [16–31]. There exist also excellent review articles [32–37] as well as a textbook [38] on this subject.

19.2 General Features of Cryogenic Calorimeters

A typical cryogenic calorimeter is shown in Fig. 19.1. It consists of three basic elements: an absorber, which confines the interaction volume, a thermometer, which is thermally well coupled to the absorber and which measures the temperature increase due to the energy loss of a particle in the absorber, and a thermal bath, which has a weak thermal link to the absorber and restores the temperature in the absorber to a defined base value. Particles interacting in the absorber material lose their energy in producing atomic and solid state excitations. These excitations produce electrons, photons (photoelectrons) and phonons. Phonons are quantized lattice vibrations which behave like particles and propagate with the speed of sound. The energy of these particles will degrade in time via electron-phonon and phonon-

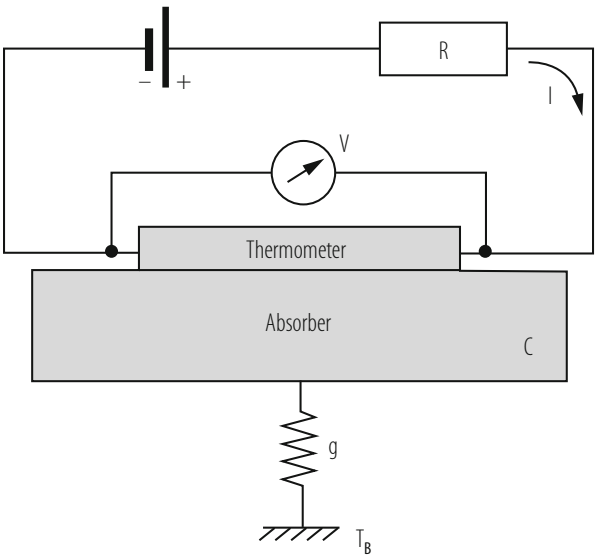


Fig. 19.1 The principle of a cryogenic calorimeter is shown

phonon interactions as well as via interactions with lattice irregularities until the system settles in thermal equilibrium. Calorimeters operating in the equilibrium mode (i.e. being sensitive to thermal phonons) offer in principle the best energy resolution, because the number of thermal phonons, with typical energies of meV, is large and the statistical fluctuations are small. For some applications thermal detectors can also be used in a non-equilibrium mode being sensitive to only high energy, so-called quasi-ballistic, phonons. These devices have the advantage of being intrinsically faster, but with energy resolutions inferior to equilibrium detectors. In calorimeters made from superconducting materials, such as superconducting tunnel junctions, the excitation energy is transformed into phonons as well as quasi-particles. These devices operate in the non-equilibrium mode, since the excitations (quasi-particles) are measured before they settle in thermal equilibrium. As described in more detail in the following paragraphs, most low temperature calorimeters differ in the way they are converting the excitation energy into a measurable signal.

Assuming that the deposited energy E of a particle in the absorber is fully thermalized the temperature rise ΔT is given by:

$$\Delta T = \frac{E}{C_{tot}}\,, \tag{19.1}$$

where $C_{tot} = cV$ is the heat capacity of an absorber with the volume V and the specific heat c. Cryogenic detectors operate at low temperatures because the heat capacity of many absorber materials becomes very small leading to an appreciable temperature rise. In addition the absorber volumes are kept as small as possible, in some cases of mm^3 or cm^3 size. Therefore they are also often called micro-calorimeters. Applying the Debye model to calculate the internal energy of the lattice vibrations (phonons), the specific heat of a dielectric crystal absorber comes out to be:

$$c_{dielectric} = \beta(\frac{T}{\theta_D})^3\,, \tag{19.2}$$

with $\beta = 1944\ \mathrm{J\ mol^{-1}\ K^{-1}}$ and θ_D the Debye temperature of the crystal. The cubic dependence on temperature demonstrates a strong decrease of the phonon specific heat at low temperatures. In a metal absorber there are two components which determine the specific heat: lattice vibrations and thermally excited conduction electrons. The specific heat of a normal conducting material at low temperatures is given by:

$$c_{metal} = \beta(\frac{T}{\theta_D})^3 + \gamma T\,, \tag{19.3}$$

with γ being a material dependent constant (Sommerfeld constant). At temperatures below 1 K the electronic specific heat dominates. Therefore the total specific heat

decreases only linearly with temperature. Another frequently used absorber is a superconductor. In this case the specific heat consists of a term due to lattice vibrations and a second term which reflects the number of thermally excited electrons across the energy gap of a superconductor Δ. The latter diminishes exponentially with temperature due to the decrease of the quasi-particle density:

$$c_{supercond.} = \beta(\frac{T}{\theta_D})^3 + a \exp\left\{(-\frac{b\Delta}{k_B T})\right\}, \tag{19.4}$$

with a and b being material constants and k_B is the Boltzmann constant. Therefore at very low temperatures the specific heat of a superconductor is dominated by lattice vibrations.

The characteristics of an ideal cryogenic calorimeter can be described by the heat capacity C of the absorber and the thermal conductivity g of the link to the heat bath with the temperature T_B. In the event of a particle losing an energy E in the absorber the temperature in the absorber will according to Eq. (19.1) rise by ΔT and then decay back to its starting temperature, which corresponds to the bath temperature T_B. The time constant of this process is given by $\tau = C/g$. The temperature rise in the absorber will change the resistance of the thermometer, which is measured by recording a voltage drop across it when passing a current I through the thermometer (Fig. 19.1). The same device can also be used to measure a continuous power input P in form of electromagnetic radiation for example. In this case the temperature rise is given by $\Delta T = P/g$. Such a device is usually referred to as a bolometer. Bolometers have a long tradition in detecting infrared radiation from astrophysical objects. They have also been used in the measurements of the cosmic microwave background radiation.

Cryogenic calorimeters can be made from many different materials including superconductors, a feature which turns out to be very useful for many applications. They can be used as targets and detectors at the same time. Due to the very small energy quanta involved they reach much higher energy resolutions than conventional ionization or solid state devices. For example, it takes only of the order of 1 meV to break a Cooper pair in a superconductor whereas a few eV are needed to create an electron-hole pair in a solid-state device. Cryogenic calorimeters are able to detect very small energy transfers, which makes them sensitive also to non-ionizing events.

The intrinsic energy resolution of a cryogenic calorimeter is limited by the thermal energy fluctuations due to the phonon exchange between the absorber and the heat sink. The mean square energy fluctuation is given by Chui et al. [43]:

$$< \Delta E^2 >= k_B T^2 C\,. \tag{19.5}$$

It is independent of the absorbed energy E, the thermal conductivity g of the heat link and the time constant τ. The above equation can intuitively be understood when assuming that the effective number of phonon modes in the detector is $N = C/k_B$, the typical mean energy of one phonon is $k_B T$ and the rms fluctuation of one phonon is one. Then the mean square energy fluctuation is $N(k_B T)^2 = k_B T^2 C$.

For a practical cryogenic calorimeter the energy resolution is therefore given to first order by

$$\Delta E_{FWHM} = 2.35\xi\sqrt{k_B T^2 C}\,, \tag{19.6}$$

where ξ is a parameter which depends on the sensitivity and noise characteristics of the thermometer and can have values between 1.2 and 2.0. The best resolution obtained so far with cryogenic calorimeters is ∼2 eV at 6 keV.

The use of superconductors as cryogenic particle detectors was motivated by the small binding energy 2Δ (order of meV) of the Cooper pairs. The breaking of a Cooper pair results in the creation of two excited electronic states the so-called quasi-particles. A particle traversing a superconductor produces quasi-particles and phonons. As long as the energy of the quasi-particles and the phonons is higher than 2Δ, they break up more Cooper pairs and continue to produce quasi-particles until their energy falls below the threshold of 2Δ. Particles which lose the energy E in an absorber produce ideally $N = E/\Delta$ quasi-particles. Thus the intrinsic energy resolution of a superconducting cryogenic calorimeter with $\Delta \approx 1$ meV is given by

$$\Delta E_{FWHM} = 2.35\sqrt{\Delta F E} \sim 2.6\,\text{eV} \tag{19.7}$$

for a 6 keV X-ray assuming a Fano factor $F = 0.2$, which is representative for most superconductors and which takes the deviation from Poisson statistics in the generation of quasi-particles into account.

For comparison, the energy resolution of a semiconductor device is typically:

$$\Delta E_{FWHM} = 2.35\sqrt{w F E} \sim 110\,\text{eV} \tag{19.8}$$

for a 6 keV X-ray, where w is the average energy necessary to produce an electron hole pair. It has a typical value of $w \approx 3$ eV. The Fano factor is $F = 0.12$ for Silicon. Because of the larger number of free charges a super-conducting device has a much better energy resolution.

In Fig. 19.2 X-ray spectra obtained with a state of the art Si(Li) solid-state device (dashed line) and a cryogenic micro-calorimeter (solid line) using a Bi absorber and an Al-Ag bilayer superconducting transition edge thermometer are compared. The micro-calorimeter has been developed at the National Institute of Standards and Technology (NIST) in Boulder (USA) [42].

19.3 Phonon Sensors

Phonons produced by a particle interaction in an absorber are far from thermal equilibrium. They must decay to lower energy phonons and become thermalized before the temperature rise ΔT can be measured. The time required to thermalize

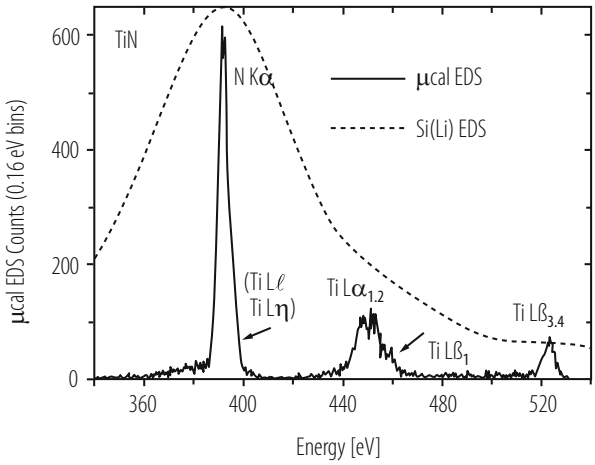


Fig. 19.2 TiN X-ray spectra obtained with a cryogenic micro-calorimeter (solid line) from the NIST group (see text) and with a state of the art Si(Li) solid-state device (dashed line) are compared. EDS stands for energy dispersive spectrometer. TiN is an interconnect and diffusion barrier material used in semiconductor industry

and the long pulse recovery time ($\tau = C/g$) limits the counting rate of thermal calorimeters to a few Hz. The most commonly used phonon sensors are resistive thermometers, like semiconducting thermistors and superconducting transition edge sensors (TES), where the resistance changes as a function of temperature. These thermometers have Johnson noise and they are dissipative, since the resistance requires power to be read out, which in turn heats the calorimeter (Joule heating). However, the very high sensitivities of these calorimeters can outweigh to a large extent these disadvantages. There are also magnetic thermometers under development, which do not have readout power dissipation.

19.3.1 Semiconducting Thermistors

A thermistor is a heavily doped semiconductor slightly below the metal insulator transition. Its conductivity at low temperatures can be described by a phonon assisted electron hopping mechanism between impurity sites. This process is also called "variable range hopping" (VRH) [44]. For temperatures between 10 mK and 4 K the resistance is expected to follow $R(T) = R_0 \exp\left\{(\frac{T_0}{T})^{\frac{1}{2}}\right\}$. This behavior is observed in doped Si and Ge thermistors. However, depending on the doping concentrations of the thermistor and the temperature range of its use, deviations from this behaviour have also been discovered. An important requirement for the fabrication of thermistors is to achieve a good doping homogeneity and reproducibility. Good uniformity of doping concentrations has been achieved either

with ion implantation or with neutron transmutation doping (NTD). In the latter case, thermal neutrons from a reactor are captured by nuclei which transform into isotopes. These can then be the donors or acceptors for the semiconductor. NTD Ge thermistors are frequently used because of their reproducibility and their uniformity in doping density. Furthermore they are easy to handle and commercially available.

It is convenient to define a dimensionless sensitivity of the thermometer:

$$\alpha \equiv \frac{d \log R}{d \log T} = \frac{T}{R} \frac{dR}{dT} . \tag{19.9}$$

The energy resolution of these devices is primarily driven by the heat capacity C of the absorber, the sensitivity of the thermometer α, the Joule heating, the Johnson noise of the load resistor and the amplifier noise. The bias current through the resistor can be optimized in such a way that it is kept high enough to provide a suitable voltage signal and low enough to minimize the Joule heating. If also the Johnson noise and the amplifier noise can be kept sufficiently low, the energy resolution of an ideal calorimeter can be described to first order by Eq. (19.6), where ξ is approximately $5(1/\alpha)^{1/2}$ [39]. For large values of α the energy resolution can be even much better than the magnitude of the thermodynamic fluctuations provided no power is dissipated by the temperature measurement of the sensor. Semiconducting thermistors have typically α values between 6 and 10, while superconducting transition edge sensors (TES) have values which are two orders of magnitude higher. A detailed description of the noise behavior and the energy resolution of cryogenic detectors can be found in [39–41].

19.3.2 Superconducting Transition Edge Sensors (TES)

A frequently used phonon sensor is the so-called transition edge sensor (TES). It consists of a very thin superconducting film or strip which is operated at a temperature in the narrow transition region between the superconducting and the normal phase, where its resistance changes between zero and its normal value R_N, as shown in Fig. 19.3a. TES sensors are usually attached to an absorber, but they can also be used as absorber and sensor at the same time. The very strong dependence of the resistance change on temperature, which can be expressed in the dimensionless parameter α of Eq. (19.9), makes the TES calorimeter sensitive to very small input energies. Superconducting strips with low T_c can have α values as high as 1000. This requires very high temperature stability. The Munich group has developed one of the first TES sensors, which was made from tungsten with a transition temperature of 15 mK [45].

The TES sensor can be operated in two different modes: the current and the voltage biased mode. In a current biased mode of operation a constant current is fed through the readout circuit as shown in Fig. 19.3b. A particle interaction in the absorber causes a temperature rise and a corresponding increase of the

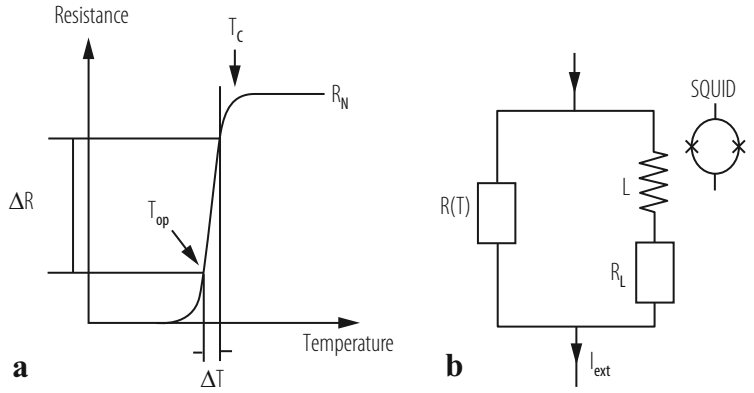


Fig. 19.3 (**a**) The temperature versus resistance diagram of a superconducting strip close to the transition temperature T_c is shown. (**b**) The dc-SQUID readout of a transition edge sensor is shown

resistance $R(T)$ of the attached TES sensor. The change of the resistance forces more current through the parallel branch of the circuit, inducing a magnetic flux change in L which is measured with high sensitivity by a superconducting quantum interference device (SQUID). However, in this mode Joule heating by the current through the sensor and small fluctuations in the bath temperature can prevent to achieve good detector performances. To solve this problem K.D. Irwin [46] has developed a so-called auto-biasing electro-thermal feedback system (ETF), which works like a thermal equivalent to an operation amplifier and keeps the temperature of the superconducting strip at a constant value within its transition region. When operating the transition edge sensor in a voltage biased mode (V_B), a temperature rise in the sensor causes an increase in its resistance and a corresponding decrease in the current through the sensor, which results in a decrease of the Joule heating ($V_B \cdot \Delta I$). The feedback uses the decrease of the Joule heating to bring the temperature of the strip back to the constant operating value. Thus the device is self-calibrating. The deposited energy in the absorber is given by $E = V_B \int \Delta I(t)dt$. It can directly be determined from the bias voltage and the integral of the current change. The use of SQUID current amplifiers allows for an easy impedance matching to the low resistance sensors and opens the possibility to multiplex the read out of large arrays of TES detectors. Another advantage of ETF is that in large pixel arrays the individual channels are self-calibrating and temperature regulated. Most important, ETF shortens the pulse duration time of TES by two orders of magnitude compared to thermistor devices allowing for higher count rates of the order of 500 Hz.

The intrinsic energy resolution for an ideal TES calorimeter is given by Eq. (19.6) with $\xi = 2(1/\alpha)^{1/2}(n/2)^{1/4}$, where n is a parameter which depends on the thermal impedance between the absorber (phonons) and the electrons in the superconducting film [46, 47]. For thin films and at low temperatures the electron-phonon decoupling dominates in the film and n is equal to 5. The best reported energy resolutions with TES devices so far are a little below 2 eV at 6 keV.

The observed transition width of TES ΔT in the presence of a typical bias current is of the order of a few mK. Large bias currents usually lead to transition broadenings due to Joule heating and self-induced magnetic fields. In order to achieve best performance of TES in terms of energy resolution or response time for certain applications, specific superconducting materials have to be selected. Both superconductors of type I and type II qualify in principle. However, the physics of the phase transition influences the noise behavior, the bias current capability and the sensitivity to magnetic field of the TES. Sensors made from high temperature superconductors have a much lower sensitivity than low temperature superconductors due to the larger gap energies Δ and heat capacities C. Thermal sensors made from strips of Al (with $T_c = 1.140\,\mathrm{K}$), Ti (0.39 K), Mo (0.92 K), W (0.012 K) and Ir (0.140 K) have been used. Ti and W sensors have been developed in early dark matter detectors [48–51]. But also other transition edge sensors, made from proximity bi-layers such as Al/Ag, Al/Cu, Ir/Au, Mo/Au, Mo/Cu, Ti/Au, or multi-layers such as Al/Ti/Au [56], have been developed to cover transition temperatures in the range between 15 and 150 mK. Although not all of these combinations are chemically stable, good detector performances have been obtained with Ir/Au bi-layers [52, 53] at transition temperatures near 30 mK. Methods to calculate bi-layer T_c can be found in [54, 55]. Another method to suppress the T_c of a superconducting film is to dope it with magnetic ions, like for example Fe (<100 ppm) [57]. However, there is a concern that the magnetic impurities may drastically increase the heat capacity of the film.

TES is also sensitive to non-thermal phonons with energies well above 2Δ. While losing energy these phonons produce quasi-particles before they thermalize. Since this process is very much faster than thermalization, signals of the order of µs can be achieved, enhancing considerably the counting rate capability of these devices as compared to thermal phonon sensors. Due to its high resolution and timing capabilities as well as versatile applications, TES sensors are currently among the most frequently used devices in calorimetric measurements. A detailed description of the performance of TES and ETF-TES can be found in [46, 47].

19.3.3 Magnetic Sensors

The magnetic properties of many materials are strongly dependent on temperature. This feature has been used to build very sensitive magnetic calorimeters applying thermal sensors made from thin paramagnetic strips, placed in a small magnetic field, which are in strong thermal contact with a suitable particle absorber. The energy deposited in the absorber leads to a temperature rise and a corresponding decrease in magnetization of the sensor. The change of magnetization is given by

$$\Delta M = \frac{dM}{dT}\frac{\Delta E}{C_{tot}} \tag{19.10}$$

with C_{tot} the total heat capacity of the thermometer and the absorber. It can be very accurately measured with a high bandwidth dc-SQUID magnetometer. The use of magnetism as thermal sensor was first developed by Buehler and Umlauf [58] and Umlauf and Buehler [59]. In these first attempts magnetic calorimeters were using the magnetization of 4f ions in dielectric host materials to measure temperature changes. Due to the weak coupling of the magnetic moments to the phonons at low temperatures these devices exhibited a too slow response time (order of seconds) for many applications. This problem was overcome by introducing sensors which use magnetic ions in metallic base material [60]. This type of device is called metallic magnetic calorimeter (MMC). In metals the relaxation times due to interactions between conduction electrons and magnetic moments are orders of magnitude faster than in dielectrics. However, the presence of conduction electrons increases the heat capacity of the sensor and leads to an enhanced interaction amongst magnetic moments. Nevertheless, very promising results were obtained with a metallic magnetic calorimeter [61]. It consisted of two thin Au disc sensors (50 µm in diameter and 25 µm thick) containing 300 ppm enriched ^{166}Er and a gold foil ($150 \times 150 \times 5\,\mu m^3$) as an X-ray absorber. The calorimeter reached an energy resolution of 3.4 eV at 6 keV, which is quite comparable to TES and thermistor calorimeters. An important property of MMC is that its inductive read out, which consists of a primary detector SQUID and a secondary SQUID amplifier, does not dissipate power into the system [61]. This feature makes MMC very attractive for many applications, in particular where large pixel arrays are of interest. The energy resolution of MMC is primarily driven by the thermal conductance between the absorber and the temperature bath and between the absorber and the sensor. For an ideal MMC the energy resolution is then given by Eq. (19.6), where ξ is approximately $\xi = 2\sqrt{2}(\tau_0/\tau)^{1/4}$ with τ_0 (typically order of µs) the relaxation time between the absorber and the sensor and τ (typically order of ms) the relaxation time between the absorber and the bath temperature [62]. It is further assumed that the heat capacities of the absorber and the sensor are approximately equal. In this case the heat capacity C in Eq. (19.6) represents the heat capacity of the absorber. MMC devices have potential applications in X-ray spectroscopy and are under further development for large pixel array cameras.

19.4 Quasiparticle Detection

The physics of superconducting detectors are based on Cooper pair breaking and quasi-particle production. Quasi-particles created by the absorption of X-rays or by the energy loss of a transient particle in a superconducting absorber can be measured with a Superconducting Tunnel Junction (STJ). The STJ device is in principle the same as the more widely known Josephson junction [63]. When biasing the STJ at a suitable voltage the tunneling current through the junction is proportional to the excess number of quasi-particles produced. To be able to measure these excess quasi-particles above the thermal background one has

to go to very low temperatures $T < 0.1\,T_c$. Arrays of STJs are also used to measure high energy, non-thermal (ballistic) phonons produced in either a dielectric or superconducting absorber. A new detector concept, called microwave kinetic inductance detector (MKID), has been developed which allows a frequency-domain approach to multiplexing and results in a dramatic simplification of the array and its associated readout electronics. Another detection scheme is based on small superheated superconducting granules (SSG) embedded in an external magnetic field. They are kept just below the phase-transition border and will change from the superconducting to the normal-conducting phase upon thermal excitation, which leads to the breaking of Cooper pairs and the penetration of the external magnetic field into the granule, causing a magnetic flux change (Ochsenfeld-Meissner effect). The flux change can be measured with an appropriate pickup coil. All these detectors are non-equilibrium devices.

19.4.1 Superconducting Tunnel Junctions (STJ)

The pioneering work of the groups of the Paul Scherrer Institute (at Villigen, Switzerland) [64] and of the Technical University Munich (Germany) [65] and their promising first results have stimulated other institutes to further develop STJs for high resolution X-ray detection. A typical STJ consists of two superconducting films S1 and S2 with a thickness of a few nm separated by a thin, 1–2 nm thick, tunnel barrier, which is usually the oxide of one of the superconductors. Because of its structure the device is frequently referred to as SIS (superconductor/insolator/superconductor) junction, also sometimes called Giaever junction. Typical junction areas are of the order of $100 \times 100\,\mu\text{m}^2$. As a quasi-particle detector, the STJ is operated with a bias voltage which is usually set to be less than Δ/e, where e is the charge of an electron. The principle processes taking place in a STJ are illustrated in Fig. 19.4, which is taken from [67]. In the event of an incident particle

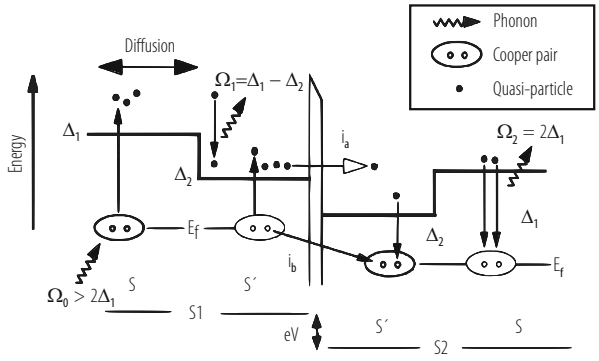


Fig. 19.4 The processes in a superconducting tunnel junction (ST) are illustrated

or X-ray interaction in film S1 the quasi-particle density is increased. This will lead to an increase of a net quasi-particle transfer from S1 to S2 and consequently to an increase of the tunneling current. However, not all quasi-particles will reach and pass through the junction barrier. Depending on the geometry and structure of the junction there will be losses. Quasi-particles can recombine to Cooper pairs radiating phonons as consequence of the relaxation process. If the phonon energy is high enough $\Omega_0 > 2\Delta_1$ to break new Cooper pairs, this process can lead to quasi-particle multiplication enhancing the signal output of the STJ. If, however, the phonon energy is below the energy threshold for breaking a Cooper pair $\Omega_0 < 2\Delta_1$ the quasi-particle will be lost and will not contribute to the signal. Quasi-particles will also be lost when they diffuse out of the overlap region of the junction films into the current leads instead of crossing the junction barrier. N. Booth [68] proposed a scheme which allows to recover some of these losses by quasi-particle trapping and in some cases quasi-particle multiplication. Quasi-particle trapping can be achieved by introducing bi-layers of superconducting materials (S(Δ_1) and S'(Δ_2)) with different gap energies $\Delta_2 < \Delta_1$ [68]. For example, an X-ray absorbed in the superconductor S produces phonons of energy $\Omega_0 > 2\Delta_1$ breaking a number of Cooper pairs. Some of the produced quasi-particles diffuse to the superconducting film S' of the STJ with a smaller gap energy Δ_2. By falling in that trap they relax to lower energies by emitting phonons, which could generate additional quasi-particles in the film S' (quasi-particle multiplication) if their energy is larger than $2\Delta_2$. However, the relaxed quasi-particles cannot diffuse back into the superconductor S because of their lower energy. They are trapped in S' and will eventually tunnel through the STJ, contributing to the signal with the tunneling current i_a. In order for quasi-particle trapping to be effective superconducting absorber materials with long quasi-particle lifetimes have to be selected (for example Al). Back tunneling, the so-called Gray effect, is also enhancing the signal [69]. In this Cooper pair mediated process a quasi-particle in film S2 recombines to form a Cooper pair at the expense of a Cooper pair in film S1. In this case the quasi-particle current i_b is also running in the direction of decreasing potential. Thus both excess quasi-particle currents i_a and i_b have the same sign. This feature allows to record signals from X-rays absorbed in either superconducting films S1 or S2 with the same sign. However, their signal shapes may not necessarily be the same due to different quasi-particle and tunneling losses in the two films. There are two other ways of electrical transport through the tunnel barrier which need to be suppressed when the STJ is used as a particle detector. One is the so-called dc Josephson current of Cooper pairs through the tunnel barrier. This current can be suppressed by applying a magnetic field of the order of a few Gauss parallel to the insulating barrier. The second is the tunnel current generated by thermally excited excess quasi-particles. The number density of these quasi-particles is decreasing with decreasing temperature according to $N_{th} \sim T^{1/2} \exp(-\Delta / k_B T)$. In order to obtain a significant signal to background ratio the operating temperature of a STJ detector should be typically lower than $0.1\, T_c$. The intrinsic energy resolution of the excess quasi-particles in a STJ device is given by Eq. (19.7), where Δ has to be replaced by ϵ, the effective energy needed to create one excited state. It turns out that $\epsilon \sim 1.7\, \Delta$ for Sn and Nb superconductors,

reflecting the fact that only a fraction of the absorbed energy is transfered into quasi-particles [70]. The number of quasi-particles generated in the STJ by an energy loss E of a particle in the superconductor is thus $N = E/\epsilon$. For a Nb superconductor with a Fano factor $F = 0.2$ and $\epsilon = 2.5\,\text{meV}$ one would expect from Eq. (19.7) an energy resolution of 4 eV at 6 keV. However, the best resolution observed so far is 12 eV at 6 keV. In order to estimate a more realistic energy resolution, quasi-particle loss and gain processes have to be taken into account. The two most important parameters driving the energy resolution of the STJ are the tunneling rate $\Gamma_t \equiv \tau_t^{-1}$ and the thermal recombination rate $\Gamma_r \equiv \tau_r^{-1}$. The temperature dependence of the thermal recombination rate is given by

$$\tau_r^{-1}(T) = \tau_0^{-1}\sqrt{\pi}\left(\frac{2\Delta}{k_B T_c}\right)^{5/2}\sqrt{\frac{T}{T_c}}\exp\left\{(-\frac{\Delta}{k_B T})\right\} \tag{19.11}$$

where τ_0 is the characteristic time of a superconductor. It has the values $\tau_0 = 2.3\,\text{ns}$ for Sn, $\tau_0 = 438\,\text{ns}$ for Al and $\tau_0 = 0.15\,\text{ns}$ for Nb [71].

The recombination rate $\Gamma_r \equiv \tau_r^{-1}$ can be minimized when operating the detector at sufficiently low temperatures, typically at $0.1\,T_c$, where the number of thermally excited quasi-particles is very small. The tunneling rate of a symmetric STJ is given by de Korte et al. [72]

$$\tau_t^{-1} = (4\,R_{norm}\cdot e^2\cdot N_0\cdot A\cdot d)^{-1}\frac{\Delta + eV_b}{\sqrt{(\Delta + eV_b)^2 - \Delta^2}} \tag{19.12}$$

where R_{norm} is the normal-conducting resistance of the junction, N_0 is the density of states of one spin at the Fermi energy, A is the junction overlap area, d the thickness of the corresponding film and V_b the bias voltage of the STJ. In practice, the tunneling time has to be shorter than the quasi-particle lifetime. For a $100\times100\,\mu\text{m}^2$ Nb-Al tunnel junctions with $R_{norm} = 15\,\text{m}\Omega$ a tunneling time of $\tau_t = 220\,\text{ns}$ and a recombination time of $\tau_r = 4.2\,\mu\text{s}$ has been measured [66]. The recombination time was determined from the decay time of the current pulse. Thus quasi-particles tunneled on average 19 times. In order to achieve even shorter tunneling times, one would have to try to further reduce R_{norm}. However, there is a fabricational limit avoiding micro-shorts in the insulator between the superconducting films. The STJ counting rate capability is determined by the pulse recovery time, which depends on the quasi-particle recombination time and can have values between several µs and up to ∼50 µs. Typical count rates of STJs are 10^4 Hz. An order of magnitude higher count rates can still be achieved, but not without losses in energy resolution. The total quasi-particle charge collected in the STJ is to first order given by

$$Q = Q_0\frac{\Gamma_t}{\Gamma_d} \tag{19.13}$$

with $Q_0 = Ne$ and $\Gamma_d = 2\Gamma_r + \Gamma_{loss}$ the total quasi-particle loss rate. The factor 2 in the recombination rate takes into account the loss of two excited electronic states and Γ_{loss} stands for all the other quasi-particle losses, mainly due to diffusion. These effects can be parametrized into an effective Fano factor which is added into the equation for the energy resolution

$$\Delta E_{FWHM} = 2.35\sqrt{\epsilon(F+G)E}\,. \tag{19.14}$$

For a symmetric tunnel junction with equal tunneling probabilities on both sides the G factor is given by $G = (1 + 1/\bar{n})$ with $\bar{n} = Q/Q_0 = \Gamma_t/\Gamma_d$ [66, 67]. It emphasizes the importance of a large tunneling rate. Still the energy resolution in Eq. (19.14) is only approximative since it neglects gain factors like quasi-particle multiplication due to relaxation phonons and loss factors due to cancellation currents, which becomes important at low bias voltage.

From Eq. (19.12) it is clear that in order to achieve a high tunnel rate the STJ detector has to be made from very thin films with a small area A. These dimensions also determine the capacitance which should be kept as small as possible in order not to degrade the detector rise time and the signal to noise ratio. For the very thin films the quantum efficiencies at X-ray energies are very low. This can be changed by separating the absorber and detector functions in fabricating devices with a larger size superconducting absorber as substrate to a STJ. Quasi-particle trapping will be achieved when choosing substrate materials with a higher energy gap with respect to the junction.

Quasi-particle trapping was first demonstrated using a Sn absorber ($975 \times 150 \times 0.25\,\mu m^3$) with an energy gap of $\Delta_{Sn} = 0.58\,meV$ and with an Al-Al_2O_3-Al STJ at each end of the absorber [65]. Quasi-particles generated by an event in the Sn absorber diffuse into the aluminum junctions, where they stay trapped because of the smaller energy gap $\Delta_{Al} = 0.18\,meV$ of Al with respect to Sn. The excess quasi-particle tunnel current was then measured with the two Al STJs. In order to prevent diffusion losses out of the Sn absorber the common contact leads to the STJs and to the Sn absorber where made from Pb, which has an even higher energy gap of $\Delta_{Pb} = 1.34\,meV$. It turned out that more than 99.6% of the quasi-particles which tried to diffuse out of the absorber where rejected at the Pb barrier and hence confined to the absorber. Currently the best energy resolution of 12 eV at 6 keV has been achieved with a single Al-Al_2O_3-Al STJ using a superconducting Pb absorber ($90 \times 90 \times 1.3\,\mu m^3$) with an absorption efficiency of ≈50% [73]. It turns out that Al is an ideal material for STJ because it allows to fabricate a very uniform layer of the tunnel barrier and has a very long quasi-particle lifetime. These are features which are essential for a high performance STJ. A very good description of the physics and applications of STJ detectors can be found in [67].

Arrays of STJs have been developed for astronomical observations and other practical applications as discussed in the chapters below. The Naples collaboration [76] produced an array of circular shaped STJs [76, 77]. This device allows the operation of STJs without external magnetic field. Position sensitive devices have been developed for reading out large pixel devices [65, 74, 78].

19.4.2 Microwave Kinetic Inductance Detector

A new detector concept, called microwave kinetic inductance detector (MKID) has been introduced with the aim to develop multi-pixel array cameras for X-ray and single photon detection [79, 80]. MKID is, like STJ and SSG, a non-equilibrium detector which is based on Cooper pair breaking and the production of quasi-particles. The basic element of the device consists of a thin superconducting film, which is part of a transmission line resonator.

The principle of detection is shown in Fig. 19.5, taken from [79]: A photon absorbed in a superconducting film will break up Cooper pairs and produce quasi-particles (a). The increase of quasiparticle density will affect the electrical conductivity and thus change the inductive surface impedance of the superconducting film, which is used as part of a transmission line resonator (b). At resonance, this will change the amplitude (c) and the transmission phase of the resonator (d).

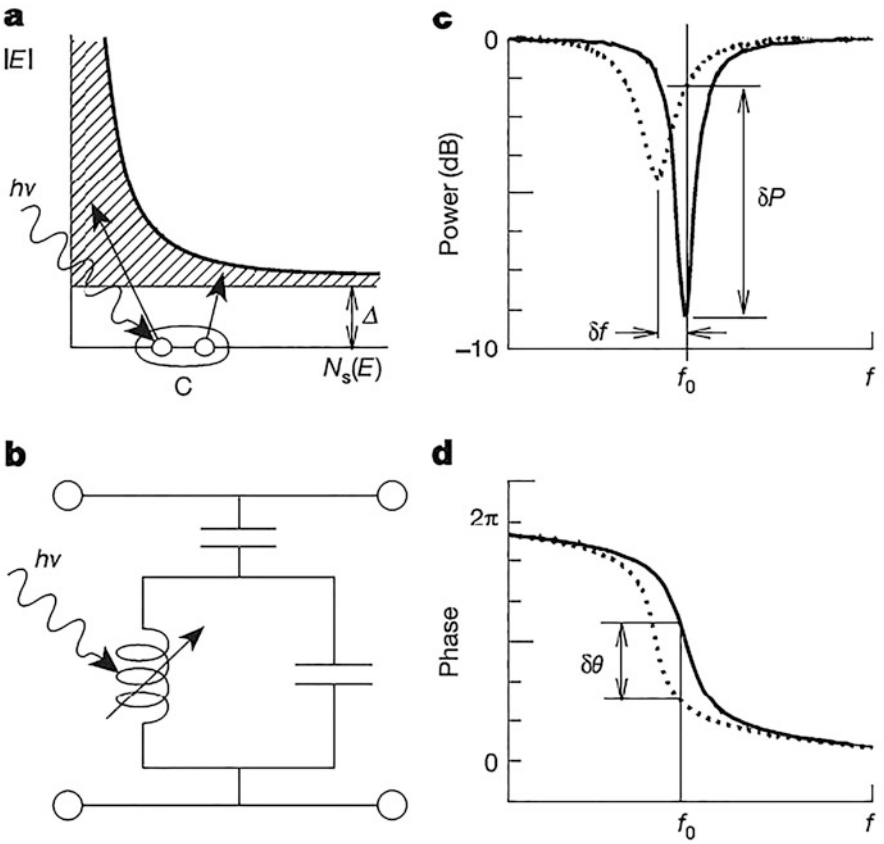


Fig. 19.5 The basic operation of a MKID (Microwave Kinetic Inductance Detector) is shown

The change in the transmission phase is proportional to the produced number of quasi-particles and thus to the photon energy. First measurements with an X-ray source yielded an energy resolution of 11 eV at 6 keV. MKID detectors find many applications where a large number of pixels are demanded. As compared to other devices multiplexing can be realized rather easy by coupling an array of many resonators with slightly different resonance frequencies to a common transmission line. A single amplifier is needed to amplify the signals from a large number of detectors. Due to its interesting features MKID is under development for many applications in ultraviolet, optical and infrared imaging [80].

19.4.3 Superheated Superconducting Granules (SSG)

Superheated superconducting granules (SSG) have been developed for X-ray imaging, transition radiation, dark matter as well as solar and reactor neutrino detection [10, 94]. A SSG detector consists of billions of small grains (typically 30 µm in diameter), diluted in a dielectric material (e.g. Teflon) with a volume filling factor of typically 10%. The detector is operated in an external magnetic field. Metastable type-1 superconductors (e.g. Sn, Zn, Al, Ta) are used, since their phase transitions from the metastable superconducting state to the normal-conducting state are sudden (in the order of 100 ns) allowing for a fast time correlation between SSG signals and those of other detectors. Its energy threshold is adjustable by setting the external magnetic field at a certain value ΔH just below the phase transition border. The phase diagram of a type-1 superconductor is schematically shown in Fig. 19.6, where H_{sh} is the superheating field, H_{sc} is the supercooling field and H_c is the critical thermodynamic field which is approximately given by $H_c(T) = H_c(0)(1 - (\frac{T}{T_c})^2)$. The region below H_{sc} is the superconducting and above H_{sh} the normal-conducting phase, while the region between the two is the so-called meta-stable phase, which is characteristic for superconductors of type-1. In order to keep the heat capacity as low as possible the SSG detector is operated at a temperature much below the critical temperature T_c at typically $T_0 \approx 100$ mK. Particles interacting in a granule produce quasi-particles. While spreading over the volume of the granule the quasi-particles are losing energy via electron-phonon interactions, thereby globally heating the granule up to a point where it may undergo a sudden phase transition (granule flip). The temperature change experienced by the granule is $\Delta T = \frac{3\Delta E}{4\pi c r^3}$, with ΔE the energy loss of the particle in the grain, c the specific heat and r the radius of the grain. The phase transition of a single grain can be detected by a pickup coil which measures the magnetic flux change $\Delta\Phi$ due to the disappearance of the Ochsenfeld-Meissner effect. In case of a single grain located in the center of the pickup coil the flux change is given by

$$\Delta\Phi = 2\pi B n \frac{r^3}{\sqrt{4R^2 + l^2}} \tag{19.15}$$

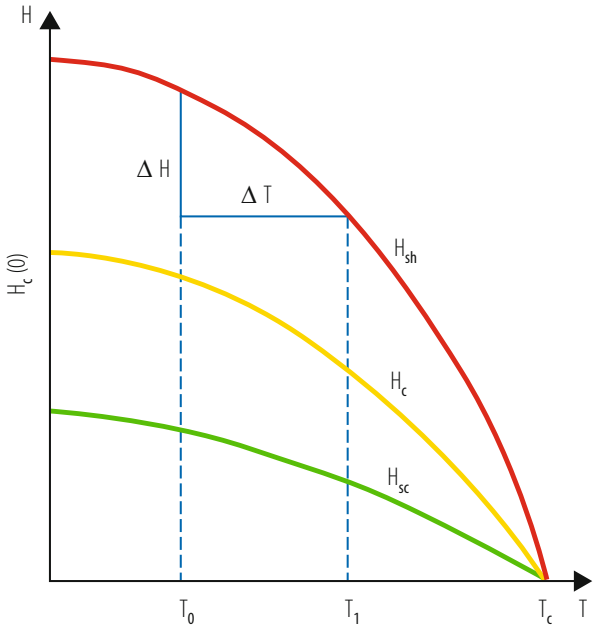


Fig. 19.6 The phase-diagram of a superconductor type1 is shown. H_{sh} is the superheating field, H_{sc} is the supercooling field and H_c is the critical thermodynamic field

with B the applied magnetic field, n the number of windings, R the radius and l the length of the pickup coil. It should be noted that one coil may contain a very large number of grains. If the flipping time τ is small compared to the characteristic time of the readout circuit ($\tau \ll 2\pi\sqrt{LC}$) the flux change induces a voltage pulse in the pick-up coil

$$V(t) = \frac{\Delta\Phi}{\omega LC} e^{-t/2RC} \sin(\omega t)\,, \quad \omega^2 = \frac{1}{LC} - \frac{1}{(2RC)^2} \tag{19.16}$$

with ω, L, R and C being parameters of the pick-up circuit. A detailed description of a readout concept using conventional pick-up coils and electronics including noise estimation is given in [81, 82]. Besides conventional readout coils more sensitive Superconducting Quantum Interference Devices (SQUID) were introduced [83–85]. The SQUID readout allows the detection of single flip signals from smaller size granules and/or the usage of larger size pickup coils. Granules of 20 µm diameter were measured in a large size prototype [85].

Small spherical grains can be produced at low cost by industry using a fine powder gas atomization technique. Since after fabrication the grains are not of a uniform diameter, they have to be sieved to select the desired size. A grain size selection within ±2 µm was achieved.

The Bern Collaboration has built and operated a dark matter SSG detector, named ORPHEUS, which consisted of 0.45 kg of spherical Sn granules with a diameter of ≈30 µm [81]. The detector was read out by 56 conventional pick-up coils, each 6.8 cm long and 1.8 cm in diameter. Each pick-up coil contained ≈80 million granules. The phase transition of each individual grain could be detected with a typical signal to noise ratio of better than 10. The principle to detect small nuclear recoil energies with SSG was successfully tested prior to the construction of the ORPHEUS detector in a neutron beam of the Paul Scherrer Institute (Villigen, Switzerland) [86]. The special cryogenics required for the ORPHEUS detector is described in [87]. The detector is located in the underground facility of the University Bern with an overburden of 70 meter water equivalent (m.w.e.). In its first phase the ORPHEUS dark matter experiment did not reach the sensitivity of other experiments employing cryogenic detectors, as described below. Further improvements on the superconducting behavior of the granules and on the local shielding are necessary.

SSG is a threshold detector. Its resolution depends on the sharpness $\delta H/H$, respectively $\delta T/T$, of the phase transition. It was found that the phase transition smearing depends on the production process of the grains. Industrially produced grains using the atomization technique exhibited a smearing of $\delta H/H \sim 20\%$. By using planar arrays of regularly spaced superheated superconducting microstructures which were produced by various sputtering and evaporation techniques the transition smearing could be reduced to about 2% [88–92]. The improvement of the phase transition smearing is one of the most important developments for future applications of SSG detectors. It looks promising that large quantities of planar arrays can be produced industrially [92].

There is a mechanism by which the energy transfered to a grain can be measured directly. If the grain is held in a temperature bath just below the H_{sc} boundary and the energy (heat) transfer to the grain is large enough to cross the meta-stable region to become normal-conducting, it will after some time cool down again to the bath temperature and become superconducting again. During this process the granule will provide a "flip" signal when crossing the H_{sh} border, and an opposite polarity "flop" signal when crossing the H_{sc} border. The elapsed time between the flip and the flop signal is a measure of the deposited energy in the grain. This effect has been demonstrated with a 11 µm Sn grain bombarded with α particles [93]. It offers the possibility to build an energy resolving and self-recovering SSG detector.

The practical realization of a large SSG detector is still very challenging. Nevertheless, the detector principle offers several unique features:

(a) The large list of suitable type-1 superconductor materials allows to optimize SSG for specific applications.
(b) Very low energy thresholds (eV) can be achieved.
(c) The inductive readout does not dissipate any power into the grains. Therefore the sensitivity of SSG is essentially determined by the grain size and the specific heat of the grain material.

(d) The sudden phase transitions are beneficial for coincident timing with other signals. Generally speaking, SSG detectors are among the most sensitive devices to detect very low energy transfers, i.e. nuclear recoils. A detailed description of SSG can be found in [10, 94, 95].

19.5 Physics with Cryogenic Detectors

19.5.1 Direct Dark Matter Detection

Among the most challenging puzzles in physics and cosmology is the existence of dark matter and dark energy. Dark matter, which was first inferred by Fritz Zwicky in 1933 [96], shows its presence by gravitational interaction with ordinary matter. It holds numerous galaxies together in large clusters and it keeps stars rotating with practically constant velocities around the centers of spiral galaxies. Dark energy, which was discovered by the Supernovae type 1a surveys in 1998 [97, 98], is driven by a repulsive force quite in contrast to the attractive gravitational force and causes the universe to expand with acceleration. The most recent information about the matter/energy content of the universe was gained from the Cosmic Microwave Background radiation (CMB) measurements by the Planck satellite [99]. According to these observations the universe contains 69.4% dark energy, 30.6% matter (including baryonic and dark matter) and 4.8% baryonic matter in form of atoms. The true nature of the dark energy and the dark matter, which fills about 95% of the universe, is still unknown. The direct detection of the dark energy, which is related to Einstein's cosmological constant, seems not to be in reach with present technologies. However, the direct detection of dark matter, if it exists in form of particles, is encouraged by the large expected particle flux which can be deduced under the following assumptions. In an isothermal dark matter halo model the velocity of particles in our galaxy is given by a Maxwell Boltzmann distribution with an average value of $< v >= 230\,\mathrm{km\,s^{-1}}$ and an upper cutoff value of $575\,\mathrm{km\,s^{-1}}$ corresponding to the escape velocity. The dark matter halo density in our solar neighborhood is estimated to be $\rho = 0.3\,\mathrm{GeV\ c^{-2}\,cm^{-3}}$. From that one expects a flux of $\Phi = \rho < v > /m_\chi \sim 7 \cdot 10^6/\mathrm{m}_\chi\,\mathrm{cm^{-2}\,s^{-1}}$ with m_χ the mass of the dark matter particle in GeV $\mathrm{c^{-2}}$. However, since neither the mass nor the interaction cross section of these particles are known one is forced to explore a very large parameter space, which requires very sensitive and efficient detection systems. The most prominent candidates for the dark matter are: massive neutrinos, WIMPs (weakly interacting massive particles) and axions. Neutrinos are among the most abundant particles in the universe, but their masses seem to be too small to contribute significantly to the missing mass. Neutrinos being relativistic at freeze out are free streaming particles, which cluster preferentially at very large scales. Therefore massive neutrinos would enhance large-scale and suppress small-scale structure formations. From hot dark matter and cold dark matter model calculations

fitting the power spectrum obtained from Large Scale Structure (LSS) surveys one obtains a value for the ratio neutrino density to matter density Ω_ν / Ω_m. From this value and Ω_m obtained from CMB an upper limit for the sum of the neutrino masses $\sum m_\nu \leq 0.234\,\mathrm{eV\ c^{-2}}$ can be derived. However, this and the results from direct neutrino mass experiments, as described below, indicate, that neutrinos have a mass to low to qualify for the dark matter. The introduction of axions was not motivated by cosmological considerations, but rather to solve the charge conjugation and Parity violation (CP) problem in Quantum Chromo-Dynamics (QCD) [100]. Nevertheless axions would be produced abundantly during the QCD phase transition in the early universe when hadrons were formed from quarks and gluons. A recent review of axion searches can be found in [101]. The most favored candidate for a WIMP is the neutralino, which is predicted by some Super Symmetric Theories (SUSY) to be the lightest stable SUSY particle. If the neutralino were to be discovered by the Large Hadron Collider (LHC) at CERN, it still would need to be confirmed as a dark matter candidate by direct detection experiments. However, up to now no sign of SUSY-particles has been observed at the LHC [102]. In the following the WIMP searches with some of the most advanced cryogenic detectors are described.

The direct detection of WIMPs is based on the measurement of nuclear recoils in elastic WIMP scattering processes. In the case of neutralinos, spin-independent coherent scatterings as well as spin-dependent scatterings are possible. The expressions for the corresponding cross sections can be found in [103, 104]. In order to obtain good detection efficiencies, devices with high sensitivity to low nuclear recoil energies (eV) are needed. WIMP detectors can be categorized in conventional and cryogenic devices. Most of the conventional WIMP detectors use NaI, Ge crystals, liquid Xenon (LXe) or liquid Argon (LAr). These devices have the advantage that large detector masses ($\sim$ton) can be employed, which makes them sensitive to annual modulations of the WIMP signal owing to the movement of the earth with respect to the dark halo rest frame. Annual modulation, if observed, would provide strong evidence for a WIMP signal, assuming it is not faked by spurious modulated background signals. However, due to quenching of the ionization signals, conventional detectors have lower nuclear recoil detection efficiencies than cryogenic devices.

Cryogenic detectors are able to measure small recoil energies with high efficiency because they measure the total deposited energy in form of ionization and heat. They can be made of many different materials, like Ge, Si, TeO_2, sapphire (Al_2O_3), LiF, $CaWO_4$ and BGO, including superconductors like Sn, Zn, Al, etc. This turns out to be an advantage for the WIMP search, since for a given WIMP mass the resulting recoil spectra are characteristically different for detectors with different materials, a feature which helps to effectively discriminate a WIMP signal against background. If the atomic mass of the detector is matched to the WIMP mass better sensitivity can be obtained due to the larger recoil energies. In comparison to conventional detectors, however, cryogenic detectors are so far rather limited in target mass ($\sim$kg).

Dark matter detectors have to be operated in deep underground laboratories in order to be screened from cosmic-ray background. In addition they need to be shielded locally against radioactivity from surrounding rocks and materials. The shielding as well as the detector itself has to be fabricated from radio-poor materials, which turns out to be rather expensive and limited in its effectiveness. Nevertheless, cryogenic detectors are capable of active background recognition, which allows to discriminate between signals from background minimum ionizing particles, i.e. Compton electrons, and signals from genuine nuclear recoils by a simultaneous but separate measurement of phonons and ionization (or photons) in each event. For the same deposited energy the ionization (or photon) signal from nuclear recoils is highly quenched compared to signals from electrons. The dual phonon-ionization detection method, which was first sugested by Sadoulet [105] and further developed by the CDMS and EDELWEISS collaborations, increases the sensitivity for WIMP detection considerably. A similar idea using scintillating crystals as absorbers and simultaneous phonon-photon detection was introduced by Gonzales-Mestres and Perret-Galix [106] and further developed by the ROSEBUD [107] and CRESST II [108] collaborations. The principle of the method is demonstrated in the scatterplot of Fig. 19.7, taken from [108]. It shows the energy equivalent of the pulse heights measured in the light detector versus those measured in the phonon detector. The scintillating $CaWO_4$ crystal absorber was irradiated with photons and electrons (using Cobalt and Strontium sources respectively) as well as with neutrons (using an Americium-Beryllium source). The photon lines visible in Fig. 19.7 were used

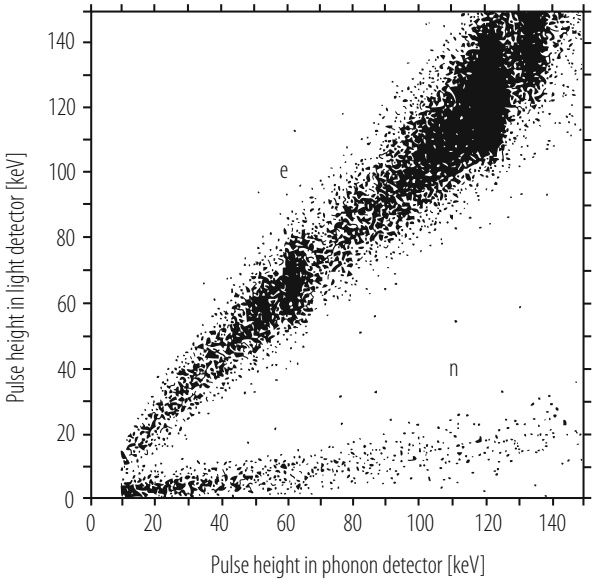


Fig. 19.7 The energy equivalent of the pulse heights measured with the light detector versus those in the phonon detector under electron, photon (e) and neutron (n) irradiation are shown

for the energy calibration in both the light and the phonon detector. The upper band in Fig. 19.7 shows electron recoils (e) and the lower band the nuclear recoils (n). Above an energy of 15 keV 99.7% of the electron recoils can be recognized and clearly distinguished from the nuclear recoils. Active background rejection was also practiced with the ORPHEUS SSG dark matter detector, since minimum ionizing particles cause many granules to flip, while WIMPs cause only one granule to flip (flip meaning a transition from superconducting to normal state) [81]. In the following some of the most sensitive cryogenic WIMP detectors in operation are described.

The CDMS experiment [109] is located at the Soudan Underground Laboratory, USA, with an overburden of 2090 meter water equivalent (m.w.e.). In an early phase of the experiment the cryogenic detectors consisted of 4 towers of 250 g Ge absorbers which where read out by NTD germanium thermistors, so called Berkeley Large Ionization and Phonon (BLIP) detectors, and two towers of 100 g Si absorbers, which where read out by TES sensors, the so-called Z-sensitive Ionization and Phonon based (ZIP) detectors. The ZIP detectors utilize tungsten aluminum Quasi-particle trapping assisted Electrothermal feedback Transition edge sensors (QET). This type of sensor covers a large area of the Si absorber with aluminum phonon collector pads, where phonons are absorbed by breaking Cooper pairs and forming quasi-particles. The quasi-particles are trapped into a meander of tungsten strips which are used as transition edge sensors. The release of the quasi-particle energy in the tungsten strips increases their resistance, which will be observed as a current change in L detected with a SQUID as indicated in Fig. 19.3b. The transition edge device is voltage biased to take advantage of the electrothermal feedback (ETF). The signal pulses of the ZIP detector have rise times of a few µs and fall times of about 50 µs. They are much faster than the signals of the BLIP detector since the ZIP detectors are sensitive to the more energetic non thermal phonons. Their sensitivity to non thermal phonons and the pad structure of the sensors at the surface of the crystal allows for a localization of the event in the x-y plane. A separate circuit collects ionization charges, which are drifted by an electric field of 3 V/cm and collected on two concentric electrodes mounted on opposite sides of the absorber. The ratio of the ionization pulse height to the phonon pulse height versus the pulse height of the phonon detector allows to discriminate nuclear from electron recoils with a rejection factor better than 10^4 and with full nuclear recoil detection efficiency above 10 keV. In order to further improve the sensitivity of the experiment CDMS II is operating at present 19 Ge (250 g each) and 11 Si (100 g each) ZIP type detectors in the Soudan Underground Laboratory at a temperature of about 40 mK. Each detector is 7.62 cm in diameter and 1 cm thick. Limits on the direct detection of WIMPs obtained with the Ge and Si detectors are published in [110] and [111] respectively. The CDMSlite (Cryogenic Dark Matter Search low ionization threshold experiment) uses the Neganov-Luke effect, which leads to an amplification of the phonon signal and allows for lower energy thresholds (56 eV) to be reached [112–114]. In this mode a large detector bias voltage is applied to amplify the phonon signals produced by drifting quasi particle charges. This opens the possibility to extend the WIMP search to masses well below 10 GeV c^{-2}.

Recent results on low mass WIMP searches for spin independent and spin dependent interactions are published in [115].

The EDELWEISS experiment [116], which is located in the Frejus tunnel (4800 m.w.e.), South of France, uses a technique similar to the CDMS BLIP detectors. It consists of 3 towers of 320 g Ge absorbers which are read out by NTD germanium thermistors. For the ionization measurement the detectors are equipped with Al electrodes which are directly sputtered on the Ge absorber crystal. For some data taking runs the EDELWEISS group used towers with amorphous Ge and Si films under the Al electrodes. More data were collected between 2005 and 2011 with the EDELWEISS II detector which contains an array of ten cryogenic Ge detectors with a mass of 400 g each [117]. As an upgrade of EDELWEISS II the collaboration developed EDELWEISS III with 36 FID (Fully Inter-Digital) detectors based on cylindrical Ge crystals with a mass of about 800 g each operating at 18 mK [118].

Early experience with the ionization measurements showed a severe limitation of the background separation capability due to insufficient charge collection of surface events. The effect can be attributed to a plasma screening of the external electric field of the electrodes. As a result surface interactions of electrons can fake nuclear recoil events. One way to solve the problem was developed by the Berkeley group by sputtering films of amorphous Si or Ge on the absorber surface before deposition of the Al electrodes [119]. Due to the modified energy band gap of the amorphous layer the charge collection efficiency was largely improved. Fast phonon detectors like ZIP allow to identify surface events by measuring the relative timing between the phonon and ionization signals as demonstrated by the CDMS experiment [120]. The surface event problem completely disappears when using dual phonon-photon detection. This method was chosen by the CRESST II collaboration.

The CRESST experiment [121] is located in the Gran Sasso Underground Laboratory (3800 m.w.e.) north of Rome, Italy. It uses scintillating $CaWO_4$ crystals as absorber material. The detector structure can hold 33 modules of absorber which can be individually mounted and dismounted. Each module weights about 300 g. The detector operates at about 10 mK. The phonon signal from the $CaWO_4$ crystal is read by a superconducting tungsten TES thermometer and the photon signal by a separate but nearby cryogenic light detector, which consists of a silicon wafer with a tungsten TES thermometer. For an effective background discrimination the light detector has to be very efficient. This was achieved by applying an electric field to the silicon crystal leading to an amplification of the thermal signal due to the Neganov-Luke effect [122]. The time constant of the emission of scintillation photons from $CaWO_4$ at mK temperatures is of the order of ms, which requires a long thermal relaxation time for the light detector. The characteristics of the background rejection power depends on the knowledge of the quenching factor, which is the reduction factor of the light output of the nuclear recoil event relative to an electron event. These quenching factors were measured by the CRESST collaboration for various recoiling nuclei in $CaWO_4$ in a separate experiment [123]. The knowledge of these quenching factors would allow in principle to identify WIMP interactions with different nuclei in the $CaWO_4$ crystal. This method seems very promising, not only for identifying the background but also the quantum

numbers of the WIMP candidates. The three types of nuclei in $CaWO_4$ together with a low nuclear recoil energy threshold of 300 eV allowed CRESST II to extend the dark matter search with high sensitivity into a mass region below 10 GeV c^{-2} [124].

Besides the dark matter search the CRESST collaboration is developing a cryogenic detector to measure coherent neutrino nucleus scattering [125]. This process is predicted by the Standard Model (SM), but has been unobserved so far. If successful, a possible application could be the real time monitoring of nearby nuclear power plants. With a small size prototype cryogenic Saphire detector with a weight of 0.5 g a recoil energy threshold of 20 eV was achieved [126]. Nevertheless, the first detection of coherent neutrino scattering was reported from the COHERENT collaboration only recently [127]. They measured neutrino-induced recoils with conventional scintillating CsI(Na) crystals with a weight of 14.5 kg. Their experiment was located in a basement under the Oak Ridge National Laboratory Spallation Neutron Source.

The experimental results are usually presented as exclusion plots, which show the WIMP-nucleon cross section versus the WIMP mass. They are derived from the expected nuclear recoil spectrum for a given set of parameters [104]:

$$\frac{dR}{dE} = \frac{\sigma_0 \rho_\chi}{2\mu^2 m_\chi} F^2(E) \int_{\upsilon_{min}}^{\upsilon_{max}} \frac{f(\upsilon)}{\upsilon} d\upsilon \tag{19.17}$$

with m_χ the mass of the WIMP, μ the reduced mass of the WIMP-nucleus system, σ_0 the total elastic cross section at zero momentum transfer, $\rho_\chi = 0.3\,\mathrm{GeV\,cm^{-3}}$ the dark matter halo density in the solar neighborhood, $F(E)$ the nuclear form factor, $f(\upsilon)$ an assumed isothermal Maxwell-Boltzmann velocity distribution of the WIMPs in the halo, $\upsilon_{min} = \sqrt{E m_N / 2\mu^2}$ the minimum velocity which contributes to the recoil energy E, and $\upsilon_{max} = 575\,\mathrm{km\,s^{-1}}$ the escape velocity from the halo. The recoil energy E is given by $E = \mu^2 \upsilon^2 (1 - \cos\theta)/m_N$, with m_N the mass of the nucleus, υ the velocity of the WIMP, and θ the scattering angle in the centre of mass system. The expected nuclear recoil spectrum for interactions with WIMPs of a given mass will then be folded with the detector response, which was obtained experimentally from calibration measurements with neutron sources or in neutron beams. From a maximum likelihood analysis, an upper limit cross section value (90% C.L.) can be extracted for several different WIMP masses. Current limits for spin-independent WIMP interactions are depicted in Fig. 19.8, taken from [128]. The Figure includes only a selection of some of the most sensitive experiments. WIMP masses below 4 GeV c^{-2} are accessible by detectors like CRESST II [124], CDMSlite [115], EDELWEISS [118] and DAMIC [129] because of their low recoil energy thresholds and/or their light absorber nuclei. For WIMP masses above 6 GeV c^{-2} the best constraints are provided by experiments like XENON1T [130], LUX [131], PANDAX II [132], XENON 100 [133] and Dark Side 50 [134], which are based on massive dual phase (liquid and gas) Xenon or Argon detectors with time projection (TPC) read out. The DAMA experiment has observed

an annual modulation signal, which they claim is satisfying the requirements of a dark matter annual modulation signal [135]. Their detector is operated in the Gran Sasso Laboratory in Italy (LNGS) and is based on highly radiopure NaI (Tl) crystal scintillators. Similar results, but less significant, were reported by the CoGeNT experiment with a cryogenic Ge detector in the Soudan Underground Laboratory (SUL) [136]. Annual modulations have not been observed by other even more sensitive experiments and the interpretation of a WIMP signal is controversial. In order to better understand the origin of the observed modulation the SABRE experiment is planning to build twin detectors one of which will be placed in the northern hemisphere at the LNGS and the other in the southern hemisphere at the Stanwell Underground Physics Laboratory (SUPL) in Australia [137]. Both detectors will be identical and based on the same target material used in the DAMA experiment.

An extraction of the spin-dependent WIMP-nucleon cross section in a model independent way is not possible, since the nuclear and the SUSY degrees of freedom do not decouple from each other. Nevertheless, when using an "odd group" model which assumes that all the nuclear spin is carried by either the protons or the neutrons, whichever are unpaired, WIMP-nucleon cross sections can be deduced. The CDMSlite experiment [115] achieved constraints for spin dependent interactions below WIMP masses of 4 GeV c^{-2} complementary to LUX, PANDAX, XENON 100 and PICASSO [138].

Several experiments are planning to extend their sensitivity to a wide range of parameter space by operating multi tonnes of target material, reducing the energy thresholds and background level until the irreducible solar, earth and atmospheric neutrino background level is reached, Fig. 19.8. The EURECA (European Underground Rare Event Calorimeter) will bring together researchers from the CRESST

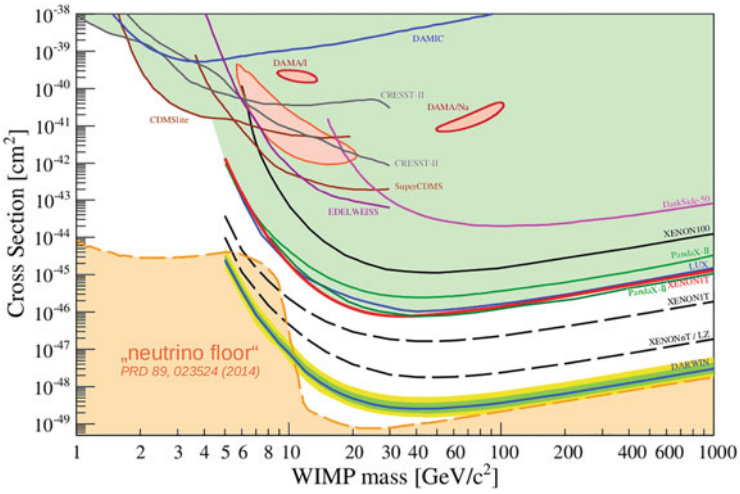


Fig. 19.8 Current limits for spin-independent WIMP interactions are shown

and EDELWEISS experiments to built a 1 ton cryogenic detector in the Modane Underground Laboratory in France [139]. SUPER CDMS will be operated in the Sudbury Neutrino Observatory (SNOLab) in Canada and is based on cryogenic Ge and Si absorber materials to increase their sensitivity for dark matter interaction cross sections to $10^{-43}\,\mathrm{cm}^2$ for masses down to $1\,\mathrm{GeV\ c}^{-2}$ [140]. The Dark Side 50 collaboration is planning to built a 23 ton dual phase liquid Argon TPC to be operated at the LNGS. The LUX-ZEPLIN (LZ) experiment is currently under construction in the Sanford Laboratory in South Dakota. It uses 10 ton of liquid XENON (dual phase) in a radio-poor double vessel cryostat [142]. The ultimate WIMP detector is proposed by the DARWIN collaboration at the LNGS [141]. It will be based on multi tonnes of liquid Xenon and will fill almost the entire parameter space for spin independent WIMP interaction cross sections down to the background level of neutrino interactions in the detector material as shown in Fig. 19.8. The ambitious project will also be sensitive to other rare interactions like solar axions, galactic axion like particles, neutrinoless double beta decay in ^{136}Xe and coherent neutrino nucleus scatterings.

19.5.2 Neutrino Mass Studies

Since the discovery of neutrino oscillations by the Kamiokande and Super-Kamiokande experiments [143] a new chapter in physics started. These findings showed that neutrinos have a mass and that there is new physics to be expected beyond the Standard Model (SM) in particle physics. Among the most pressing questions remain the absolute values of the neutrino masses, since from oscillation experiments only mass differences can be obtained [144], and the Dirac or Majorana type character of the neutrino. The main streams in this field focus upon the search for the neutrinoless double beta decay and the endpoint energy spectrum of beta active nuclei. Cryogenic detectors are particularly well suited for this type of research since they provide excellent energy resolutions, an effective background discrimination and a large choice of candidate nuclei.

19.5.2.1 Neutrinoless Double Beta Decay

Double beta decay was first suggested in 1935 by Maria Goeppert Mayer [145]. It is the spontaneous transition from a nucleus (A,Z) to its isobar (A,Z+2). This transition can proceed in two ways: (A,Z)⇒(A,Z+2) + 2 e^- + 2 $\bar{\nu}_e$ or (A,Z)⇒(A,Z+2) + 2 e^-. In the first channel, where two electrons and two antineutrinos are emitted, the lepton number is conserved. It is the second channel, the neutrinoless double beta decay ($0\nu\beta\beta$), where the lepton number is violated. In this case, with no neutrino in the final state, the energy spectrum of the decay would show in a peak which represents the energy sum of the two electrons. The experimental observation of this process would imply that neutrinos are Majorana particles, meaning that the

neutrino is not distinguishable from its antiparticle and that it has a non-vanishing mass. From the measured decay rate $(1/T^{0\nu}_{1/2})$ one can derive in principle its effective mass $< m_\nu >$ or a lower limit of it:

$$(1/T^{0\nu}_{1/2}) = G_{0\nu}(E_0, Z) \mid M_{0\nu} \mid^2 < m_\nu >^2 \tag{19.18}$$

where $G_{0\nu}(E_0, Z)$ is an accurately calculable phase space function and $M_{0\nu}$ is the nuclear matrix element, which is not very well known [146]. The calculated values of $M_{0\nu}$ can vary by factors up to two. Consequently the search for $0\nu\beta\beta$ should be made with several different nuclei in order to confirm an eventual discovery of this important process.

The Milano group has developed an experiment with the name CUORICINO to search for the neutrinoless double beta decay of ^{130}Te. The experiment is located in the Gran Sasso Underground Laboratory. The detector consists of an array of 62 TeO_2 crystals with the dimensions $5 \times 5 \times 5\,cm^3$ (44 crystals) and $3 \times 3 \times 3\,cm^3$ (18 crystals) and a total mass of 40.7 kg. The crystals are cooled to ~8 mK and attached to Ge NTD thermistors for phonon detection. Among other possible nuclear candidates (like for example $^{48}CaF_2$, ^{76}Ge, $^{100}MoPbO_4$, $^{116}CdWO_4$, $^{150}NdF_3$, $^{150}NdGaO_3$), $^{130}TeO_2$ was chosen because of its high transition energy of 2528.8 ± 1.3 keV and its large isotopic abundance of 33.8%. Published first results of the CUORICINO experiment [147] show no evidence for the $0\nu\beta\beta$ decay, but they set a lower limit on the half lifetime $T^{0\nu}_{1/2} \geq 1.8 \cdot 10^{24}$ yr (90% C.L.) corresponding to $< m_\nu > \leq$ 0.2 to 1.1 eV (depending on nuclear matrix elements). In a next step the collaboration developed CUORE-0 as prototype for a larger detector CUORE. Its basic components consist of 52 TeO_2 crystals with dimensions $5 \times 5 \times 5\,cm^3$ and a total weight of 39 kg corresponding to 10.9 kg ^{130}Te. CUORE-0 was operated in the CUORICINO cryostat at 12 mK. The data taken from 2013 to 2015 show no evidence for a neutrinoless double beta signal. Combined with the CUORICINO results a limit on the half lifetime $T^{0\nu}_{1/2} \geq 4 \cdot 10^{24}$ yr (90% C.L.) corresponding to $< m_\nu > \leq$ 270 to 760 meV (depending on nuclear matrix elements) was achieved [148]. CUORE, contains 19 CUORE-0 type towers with 988 TeO_2 crystals of a total mass of 741 kg corresponding to 206 kg of ^{130}Te [149, 150]. The array will be cooled in a large cryostat to 10 mK. It started commissioning early 2017 and aims for a sensitivity to reach limits of $T^{0\nu}_{1/2} \geq 9 \cdot 10^{25}$ yr in 5 years running time [150]. For the future the CUPID collaboration plans to develope a tonne-scale cryogenic detector which will be based on the experience gained with the CUORE experiment [151].

Several experiments investigated other nuclei and set stringent upper limits on the decay rates, for example: KamLand-Zen in ^{136}Xe [152], EXO-200 in ^{136}Xe [153], GERDA in ^{76}Ge [154], NEMO-3 in ^{100}Mo [155]. So far no neutrinoless double beta signal was seen. Currently a limit on the half lifetime $T^{0\nu}_{1/2} \geq 1.07 \cdot 10^{26}$ yr (90% C.L.) corresponding to $< m_\nu > \leq$ 60 to 165 meV (depending on nuclear matrix elements) was achieved by the KamLand-Zen experiment. An ambitious alternative approach in looking for Majorana versus Dirac type neutrinos is proposed by the

PTOLEMY experiment in studying the interaction of cosmic relic neutrinos with Tritium [156].

19.5.2.2 Direct Neutrino Mass Measurements

So far the best upper limit for the electron neutrino mass of 2.2 eV was obtained from the electron spectroscopy of the tritium decay $^3\mathrm{H} \Rightarrow {}^3\mathrm{He} + e^- + \bar{\nu}_e$, with a transition energy of 18.6 keV, by the Mainz and the Troitsk experiments [157]. In the near future the KATRIN experiment, which measures the same decay spectrum with a much improved electron spectrometer, will be in operation aiming for a neutrino mass sensitivity down to 0.2 eV [158].

One of the problems with experiments based on a spectroscopic measurement of the emitted electrons is that they yield negative values for the square of the neutrino mass when fitting the electron energy spectrum. This is mainly due to final state interactions (like tritium decays into excited atomic levels of ^{3}He), which lead to deviations from the expected energy spectrum of the electron. Low temperature calorimeters provide an alternative approach, since they measure the total energy including final state interactions, such as the de-excitation energy of excited atomic levels. However, in order to reach high sensitivity for low neutrino masses the detector has to have an excellent energy resolution and enough counting rate statistics at the beta endpoint energy. The Genoa group [159] pioneered this approach and studied the beta decay of $^{187}\mathrm{Re} \Rightarrow {}^{187}\mathrm{Os} + e^- + \bar{\nu}_e$ with a cryogenic micro-calorimeter. Their detector was a rhenium single crystal (2 mg) coupled to a Ge NTD thermistor. Rhenium is a super-conductor with a critical temperature of 1.7 K. Natural Re contains 62.8% of ^{187}Re with an endpoint energy of about 2.6 keV. The operating temperature of the detector was $T = 90$ mK. In their first attempt they obtained precise values for the beta endpoint energy and the half life of the ^{187}Re beta decay and were able to obtain an upper limit of the electron neutrino mass of 19 eV (90% CL) or 25 eV (95% CL) [160]. Following this approach the Milan group [161] has built an array of ten thermal detectors for a ^{187}Re neutrino mass experiment. The detectors were made from $AgReO_4$ crystals with masses between 250 and 350 µg. The crystals were coupled to Si implanted thermistors. Their average energy resolution (FWHM) at the beta endpoint was 28.3 eV, which was constantly monitored by means of fluorescence X-rays. The natural fraction of ^{187}Re in $AgReO_4$ yields a decay rate of $5.4 \cdot 10^{-4}$ Hz/µg. From a fit to the Curie plot of the ^{187}Re decay they obtained an upper limit for $m_{\bar{\nu}_e} \leq 15$ eV. Their measured value for the beta endpoint energy is 24653.3 ± 2.1 eV and for the half live is $(43.2 \pm 0.3) \cdot 10^9$ yr. A higher sensitivity to low neutrino masses may be achievable in the future, provided that the energy resolution and the statistics at the beta endpoint energy can be improved significantly. The latter may raise a problem for thermal phonon detectors, since their signals are rather slow and therefore limit the counting rate capability to several Hz.

With their Rhenium cryogenic micro-calorimeters the Genoa group [162, 163] and the Milano collaboration [164] were also able to measure interactions between

the emitted beta particle and its local environment, known as beta environmental fine structure (BEFS). The BESF signal originates from the interference of the outgoing beta electron wave and the reflected wave from the atoms in the neighbourhood. BEFS is similar to the well known Extended X-ray Absorption Fine Structure (EXAFS) method. Their results demonstrated that cryogenic micro-calorimeters may also offer complementary new ways for material sciences to study molecular and crystalline structures.

Currently several groups MARE [165, 166], ECHo [167], HOLMES [168], NUMECS [169], are investigating the possibility to measure the electron neutrino mass from the Electron Capture (EC) decay spectrum of Holmium (^{163}Ho) using cryogenic micro-calorimeters. This approach was originally suggested by A. de Rujula and M. Lusignoli in 1982 [170]. ^{163}Ho decays via EC into ^{163}Dy with a half life of 4570 years and a decay energy of 2.833 keV. It does not occur naturally and it is not commercially available. It has to be produced by neutron or proton irradiation. After purification the ^{163}Ho atoms have to be implanted into a suitable absorber material of the micro-calorimeter. In order to reach a neutrino mass sensitivity in the sub-eV region a total ^{163}Ho activity of several MBq is required. Since the activity of a single micro-calorimeter should not exceed 100 Bq the total ^{163}Ho activity has to be distributed over a large number of pixels (10^5). The groups are devoting much effort in developing micro-calorimeters with energy and time resolutions of the order of 1 eV and 1 µs respectively. Various thermal sensors, like TES, MMC and MKID, are considered. Multiplexing schemes have still to be invented to be able to read out the enormous number of pixels.

As already mentioned above, an upper limit for all neutrino masses of $\sum m_\nu \leq 0.234\,\mathrm{eV\,c^{-2}}$ was derived from cosmology. It will still take some efforts to reach or go below these limits in the near future with direct mass measurements. A review of direct neutrino mass searches can be found in [171].

19.5.3 Astrophysics

Modern astrophysics addresses a large list of topics: Formation of galaxies and galaxy clusters, the composition of the intergalactic medium, formation and evolution of black holes and their role in galaxy formation, matter under extreme conditions (matter in gravitational fields near black holes, matter inside neutron stars), supernovae remnants, accretion powered systems with white dwarfs, interstellar plasmas and cosmic microwave background radiation (CMB). The investigation of these topics requires optical instruments with broad band capability, high spectral resolving power, efficient photon counting and large area imaging properties. The radiation received from astrophysical objects spans from microwaves, in the case of CMB, to high energy gamma rays. The subjects discussed here can be divided into three categories, X-ray, optical/ultraviolet (O/UV) and CMB observations. In order to avoid the absorptive power of the earth atmosphere many of the instruments are operated in orbiting observatories, in sounding rockets or in balloons. Progress in

this rapidly growing field of science is constantly asking for new instrumentation and new technologies. Cryogenic detectors are playing a key role in these developments providing very broad-band, imaging spectrometers with high resolving power. They also feature high quantum efficiencies, single photon detection and timing capabilities. The observation of large-scale objects, however, needs spatial-spectral imaging devices with a wide field of view requiring cryogenic detectors to be produced in large pixel arrays. The fabrication and the readout of these arrays remains still a big challenge.

19.5.3.1 X-Ray Astrophysics

The orbital X-ray observatories Chandra and XMM-Newton contained CCD cameras for large field imaging and dispersive spectrometers for narrow field high spectral resolution. Cryogenic devices are able to combine both features in one instrument. Although they have not yet reached the imaging potential of the 2.5 megapixel CCD camera on XMM-Newton and the resolving power $E/\Delta E_{FWHM} = 1000$ at $E = 1\,\mathrm{keV}$ of the dispersive spectrometer on Chandra, their capabilities are in many ways complementary. For example, the resolving power of cryogenic devices increases with increasing energy and is above 2 keV better than the resolving power of dispersive and grating spectrometers, which decreases with increasing energy. Since the cryogenic pixel array provides a complete spectral image of the source at the focal plane its resolving power is independent of the source size. Cryogenic detectors also provide precise timing information for each photon allowing to observe rapidly varying sources such as pulsars, etc. They cover a wide range of photon energies (0.05–10 keV) with a quantum efficiency of nearly 100%, which is 5 times better than the quantum efficiency (20%) of CCDs.

The first space-borne cryogenic X-ray Quantum Calorimeter (XQC), a collaboration between the Universities of Wiskonsin, Maryland and the NASA Goddard Space Flight Center, was flown three times on a sounding rocket starting in 1995 [172]. The rockets achieved an altitude of 240 km providing 240 s observation above 165 km per flight. The XQC was equipped with a 2×18 micro-calorimeter array consisting of HgTe X-ray absorbers and doped silicon thermistors yielding an energy resolution of 9 eV across the spectral band. The pixel size was $1\,\mathrm{mm}^2$. The micro-calorimeters were operated at 60 mK. It is interesting to note that, when recovering the payload after each flight, the dewar still contained some liquid helium. The purpose of the mission was to study the soft X-ray emission in the band of 0.03–1 keV. The physics of the diffuse interstellar X-ray emission is not very well understood. It seems that a large component is due to collisional excitations of particles in an interstellar gas with temperatures of a few 10^6 K. A detailed spectral analysis would allow to determine the physical state and the composition of the gas. Since the interstellar gas occupies a large fraction of the volume within the galactic disk, it plays a major role in the formation of stars and the evolution of the galaxy. The results of this experiment and their implications are discussed in [172]. As a next step the Japan-USA collaboration has put an X-ray spectrometer

(XRS) on board of the Astro-E2 X-ray Suzaku satellite which was launched in July 2005. This instrument is equipped with 32 pixels of micro-calorimeters (HgTe) and semiconducting thermistors and is an improvement over the XQC spectrometer in terms of fabrication techniques, thermal noise, energy resolution of 7 eV across the operating band of 0.03–10 keV and observation time [173, 174]. The pixels are 0.624 mm^2 and arranged in a 6×6 array giving a field of view of 2.9×2.9 arcmin. The observatory is looking at the interstellar medium in our and neighboring galaxies as well as at supernovae remnants. The investigations include super massive black holes and the clocking of their spin rate.

The next step is to develop cryogenic detectors with increased pixel numbers (30×30) and energy resolutions of 2 eV, which would be able to replace dispersive spectrometers in future experiments. Superconducting micro-calorimeters with TES sensors have the potential to reach energy resolutions of 2 eV and are likely to replace semiconducting thermistors. One of the problems with cryogenic detectors is that they do not scale as well as CCDs, which are able to clock the charges from the center of the arrays to the edge using a serial read out. Cryogenic detectors rely on individual readout of each pixel. Another problem is the power dissipation in large arrays. To solve some of these problems, dissipation-free metallic magnetic calorimeters (MMC) and microwave kinetic inductance detectors (MKID) [80] are among the considered possibilities. Details of these new developments can be found in the proceedings of LTD. Large cryogenic detector arrays are planned for ATHENA (Advanced Telescope for High Energy Astrophysics), a future X-ray telescope of the European Space Agency. It is designed to investigate the formation and evolution of large scale galaxy clusters and the formation and grows of super massive black holes. The launch of ATHENA is planned for 2028.

19.5.3.2 Optical/UV and CMB Astrophysics

Since the first optical photon detection with STJ's in 1993 [175] and TES's in 1998 [176] a new detection concept was introduced in the field of Optical/UV astrophysics. Further developments demonstrated the potential of these single photon detection devices to combine spectral resolution, time resolution and imaging in a broad frequency band (near infrared to ultraviolet) with high quantum efficiency. The principle of these spectrophotometers is based on the fact that for a superconductor with a gap energy of typically 1 meV an optical photon of 1 eV represents a large amount of energy. Thus a photon impinging on a superconductor like for example Ta creates a large amount of quasi-particles leading to measurable tunnel current across a voltage biased junction. A first cryogenic camera S-Cam1 with a 6×6 array of Ta STJs (with a pixel size of $25 \times 25\,\mu m^2$) was developed by ESTEC/ESA and 1999 installed in the 4 m William Herschel telescope on La Palma (Spain) [75]. For a first proof of principle of this new technique the telescope was directed towards the Crab pulsar with an already known periodicity of 33 ms. The photon timing information was recorded with a 5 µs accuracy with respect to the GPS timing signals. Following the success of the demonstrator model S-Cam1

and of the improved model S-Cam2 a new camera S-Cam3 consisting of a 10×12 Ta STJ pixel array (with pixel dimension $33 \times 33\,\mu m^2$) was installed at the ESA 1m Optical Ground Station Telescope in Tenerife (Spain). The STJ structure is 100 nm Ta/30 nm Al//AlOx//30 nm Al/100 nm Ta. The camera covers a wavelength range of 340–740 nm with a wavelength resolution of 35 nm at $\lambda = 500$ nm and has a pulse decay time of 21 µs [177]. The Stanford-NIST (National Institute of Standards and Technology) collaboration has also developed a camera with an 8×8 pixel array based on tungsten TES sensors on a Si substrate [178]. Each pixel has a sensitive area of $24 \times 24\,\mu m^2$ and the array has a $36 \times 36\,\mu m$ center to center spacing. In order to improve the array fill factor a reflection mask is positioned over the inter-pixel gaps. For both STJ and TES spectro-photometers thermal infrared (IR) background radiation, which increases rapidly with wavelength above 2 µm, is of concern. Special IR blocking filters have to be employed in order to extend the wavelengths range of photons from 0.3 µm out to 1.7 µm. A 4 pixel prototype of the Stanford-NIST TES instrument was already mounted at the 2.7 m Smith Telescope at McDonald Observatory (U.S.A) and observed a number of sources including spin powered pulsars and accreting white dwarf systems. The Crab pulsar served as a source to calibrate and tune the system. The obtained data are published in [179]. Already these first results have shown that STJ or TES based spectrophotometers are in principal very promising instruments to study fast time variable sources like pulsars and black hole binaries as well as faint objects, like galaxies in their state of formation. However, in order to extend the observations from point sources to extended objects much larger pixel arrays are required. Future developments concentrate on a suitable multiplexing system in order to increase the number of pixels, which are presently limited by the wiring on the chip and the size of the readout electronics. SQUID multiplexing readout systems [180–182] as well as Distributed Read-out Imaging Devices (DROID) [74, 183], in which a single absorber strip is connected to two separate STJs on either side to provide imaging capabilties from the ratio of the two signal pulses, are under study. However, these devices are slower than small pixel devices and can handle only lower count rates. A much faster device is the superconducting MKID detector [80], which allows a simple frequency-domain approach to multiplexing and profits from the rapid advances in wireless communication electronics. A more detailed review can be found in [184, 185]. A camera, ACRONS, for optical and near infrared spectroscopy has been developed. The camera contains a 2024 pixel array of cryogenic MIKD detectors. The device is able to detect individual photons with a time resolution of 2 µs and simultaneous energy information [80]. The instrument has been used for optical observations of the Crab pulsar [186].

Large cryogenic pixel antennas have been developed for ground-based and space-borne CMB polarization measurements. These devices aim to be sensitive to the detection of the so-called primordial E- and B-modes, which would appear as curling patterns in the polarization measurements. E-modes arise from the density perturbations while B- modes are created by gravitational waves in the early universe. The two modes are distinguishable through their characteristic patterns. However, B-mode signals are expected to be an order of magnitude weaker than E-

modes. Nevertheless, B-modes are of particular interest since they would provide revealing insight into the inflationary scenario of the early universe signaling the effect of primordial gravity waves. There are several instruments, which use TES based cryogenic bolometers, in operation: EBEX [187] and SPIDER [188] are balloon-borne experiments, overflying the Antarktis. POLARBEAR [189] is an instrument, which is coupled to the HUAN TRAN Teleskope at the James Ax Observatory in Chile. BICEP2 and the Keck Array [190] are located at the Amundsen-Scott South Pole Station. The PLANCK Satellite [191] carried 48 cryogenic bolometers operating at 100 mK in outer space. In March 2014 the BICEP2 collaboration reported the detection of B-modes [190]. However the measurement was received with some skepticism and David Spergel argued that the observation could be the result of light scatterings off dust in our galaxy. In September 2014 the PLANCK team [191] concluded that their very accurate measurement of the dust is consistent with the signal reported by BICEP2. In 2015 a joint analysis of BICEP2 and PLANCK was published concluding that the signal could be entirely attributed to the dust in our galaxy [192].

19.6 Applications

Electron probe X-ray microanalysis (EPMA) is one of the most powerful methods applied in material sciences. It is based on the excitation of characteristic X-rays of target materials by high current electron beams in the energy range of several keV. EPMA finds its application in the analysis of contaminant particles and defects in semiconductor device production as well as in the failure analysis of mechanical parts. It is often used in the X-ray analysis of chemical shifts which are caused by changes in the electron binding due to chemical bonding as well as in many other sciences (material, geology, biology and ecology). The X-rays are conventionally measured by semiconducting detectors (Si-EDS), which are used as energy dispersive spectrometers covering a wide range of the X-ray spectrum, and/or by wavelength dispersive spectrometers (WDS). Both devices have complementary features. WDS, based on Bragg diffraction spectrometry, has a typical energy resolution of 2–15 eV (FWHM) over a large X-ray range. However, the diffraction limits the bandwidth of the X-rays through the spectrometer as well as the target size, which acts as a point source, and makes serial measurements necessary, which is rather time consuming. Contrary, the Si-EDS measures the entire X-ray spectrum from every location of the target simultaneously, but with a typical energy resolution of 130 eV. Si-EDS is therefore optimally suited for a quick but more qualitative analysis. Cryogenic micro-calorimeter EDS provides the ideal combination of the high resolution WDS and the broadband features of the energy dispersive EDS. The application of cryogenic detectors for EPMA was first introduced by Lesyna et al. [193]. The NIST group developed a prototype TES based micro-calorimeter which is suitable for industrial applications [42, 194, 195]. It consists of a Bi absorber and a Al-Ag or Cu-Mo bi-layer TES sensor. It covers

an area of $0.4 \times 0.4\,\mathrm{mm}^2$ and yields an energy resolution of 2 eV at 1.5 keV and 4.5 eV at 6 keV. The detector is cooled to 100 mK by a compact adiabatic demagnetization refrigerator and is mounted on a scanning electron microscope. Cryogenic refrigerators for this and various other applications are commercially available [196]. In spite of its excellent resolving power (see Fig. 19.2) the micro-calorimeter EDS has still two shortcomings. It has a limited counting rate capability (1 kHz compared to 100 kHz WDS and 25 kHz Si-EDS) and a small effective detector surface. To compensate for the latter the NIST group developed an X-ray focusing device using poly-capillary optics. The device consists of many fused tapered glass capillaries which focus the X-rays by means of internal reflection onto the micro-calorimeter increasing its effective area. Another solution under study is a multiplexed micro-calorimeter array with a possible loss in resolution due to the variability of individual detectors [182]. Nevertheless, TES based cryogenic micro-calorimeters EDS have already demonstrated major advances of EPMA in scientific and industrial applications.

Time of flight mass spectrometry (ToF-MS) of biological molecules using cryogenic detectors was first introduced by D. Twerenbold [197]. The main advantage of cryogenic calorimeters over traditionally employed micro-channel plates (MP) is that the former are recording the total kinetic energy of an accelerated molecule with high efficiency, independent of its mass, while the efficiency of the latter is decreasing with increasing mass due to the reduction of the ionization signal. ToF-MS equipped with MP lose rapidly in sensitivity for masses above 20 kDa (proton masses). The disadvantages of cryogenic detectors are: first, they cover only a rather small area of $\sim 1\,\mathrm{mm}^2$ while a MP with $4\,\mathrm{cm}^2$ will cover most of the beam spot size of a spectrometer; second, the timing signals of the cryogenic calorimeters are in the range of µs and therefore much slower than ns signals from MP, which degrades the flight time measurements and thus the accuracy of the molecular mass measurements. A good review of early developments can be found in [198]. After the early prototype experiences made with STJs [199, 200] and NIS (Normal-conductor/Isolator/Super-conductor) tunnel junctions [201], which provided only very small impact areas, super-conducting phase transition thermometers (SPT) with better time of flight resolutions and larger impact areas of $3 \times 3\,\mathrm{mm}^2$ were developed [202–204]. These devices consist of thin super-conducting Nb meanders, or super-conducting films in thermal contact with an absorber, which are current biased and locally driven to normal conducting upon impact of an ion. A voltage amplifier is used to measure the signal pulse.

Cryogenic detectors as high resolution γ-ray, α and neutron spectrometers also found applications in nuclear material analysis, as broad band micro-calorimeters in Electron Beam Ion Traps (EBITs) and in synchrotrons for fluorescence-detected X-ray absorption spectroscopy (XAS) [202]. They are also employed in nuclear and heavy ion physics [205].

19.7 Summary

Cryogenic detectors have been developed to explore new frontiers in astro and particle physics. Their main advantages over more conventional devices are their superior energy resolution and their sensitivity to very low energy transfers. However, most thermal detectors operate at mK temperatures requiring complex refrigeration systems and they have limited counting rate capabilities (1 Hz–1 kHz). Today the most frequently used thermometers for calorimeters operating in near equilibrium mode are doped semiconductors (thermistors), superconducting transition edge sensors (TES) and metallic paramagnets (MMC). Because they are easy to handle and commercially available thermistors are quite popular. They have, however, the disadvantage of having to deal with Joule heating introduced by their readout circuit. The most advanced technology is provided by the TES sensors in connection with an auto-biasing electrothermal feedback system. This system reduces the effect of Joule heating, stabilizes the operating temperature and is self-calibrating, which turns out to be advantageous also for the operation of large detector arrays. The main advantage of MMCs is their magnetic inductive readout, which does not dissipate power into the system. This feature makes MMC attractive for applications, where large detector arrays are required. Non-equilibrium detectors like superconducting tunnel junctions (STJ), superheated superconducting granules (SSG) and microwave kinetic inductance devices (MKID) are based on the production and detection of quasi-particles as a result of Cooper pair breaking in the superconductor. These devices are intrinsically faster, providing higher rate capabilities (10 kHz and more) and good timing properties suitable for coincident measurements with external detectors. Because of their sensitivity to low energy photons arrays of STJs are frequently employed in infrared and optical telescopes, but also efficiently used in x-ray spectroscopy. SSG detectors with inductive readout have the potential to reach very low energy thresholds (order of several eV) which would be advantageous for various applications like for example in neutrino physics (coherent neutrino scattering, etc.). But the practical realization of SSG detectors is still very challenging. MKIDs provide an elegant way to readout large detector arrays by coupling an array of many resonators with slightly different resonance frequencies to a common transmission line with a single signal amplifier. Due to this feature they are very suited for the future instrumentation of astrophysical observatories and other applications. Cryogenic calorimeters with a large detector mass for dark matter searches and neutrino physics as well as large detector arrays for astrophysical measurements and other practical applications are under intense developments. Despite the enormous progress made in the past their fabrication and readout remain still a challenge.

References

1. S.P. Langley, Proc. Am. Acad. Arts Sci. **16** (1881) 342.
2. P. Curie, A. Laborde, Compt.Rend. Hebd. Seances Acad. Sci. Paris **136** (1903) 673–675.
3. C.D. Ellis, A. Wooster, Proc. R. Soc. **117** (1927) 109–123.
4. W. Orthmann, Z. Phys. **60** (1930) 10; and L. Meitner, W. Orthmann, Z. Phy. **60** (1930) 143.
5. F. Simon, Nature **135** (1935) 763.
6. D.H. Andrews, R.D. Fowler, M.C. Williams, Phys.Rev. **76** (1949) 154.
7. G.H. Wood, B.L. White, Appl. Phys. Lett. **15** (1969) 237; and G.H. Wood, B.L. White, Can. J. Phys. **51** (1973) 2032.
8. H. Bernas et al., Phys. Lett. A **24** (1967) 721.
9. A. Drukier, C. Vallette, Nucl. Instrum. Meth. **105** (1972) 285.
10. A. Drukier, L. Stodolsky, Phys. Rev. D **30** (1984) 2295.
11. N. Coron, G. Dambier, J. Leblanc, in: *Infrared detector techniques for Space Research*, V. Manno, J. Ring (eds.), Reidel Dordrecht (1972), pp. 121–131.
12. T.O. Niinikoski, F. Udo, CERN NP Report 74–6 (1974).
13. E. Fiorini, T.O. Niinikoski, Nucl. Instrum. Meth. **224** (1984) 83.
14. D. McCammon, S.H. Moseley, J.C. Mather, R. Mushotzky, J. Appl. Physics **56**(5) (1984) 1263.
15. N. Coron et al., Nature **314** (1985) 75–76.
16. K. Pretzl, N. Schmitz, L. Stodolsky (eds.), Low-Temperature Detectors for Neutrinos and Dark Matter LTD1, Schloss Ringberg, Germany, Springer-Verlag (1987).
17. L. Gonzalez-Mestres, D. Perret-Gallix (eds.), Low-Temperature Detectors for Neutrinos and Dark Matter LTD2, Gif-sur-Yvette, France, Ed. Frontieres (1988).
18. L. Brogiato, D.V. Camin, E. Fiorini (eds.), Low-Temperature Detectors for Neutrinos and Dark Matter LTD3, Gif-sur-Yvette, France, Ed. Frontieres (1989).
19. N.E. Booth, G.L. Salmon (eds.), Low Temperature Detectors for Neutrinos and Dark Matter LTD4, Gif-sur-Yvette, France, Ed. Frontieres (1992).
20. S.E. Labov, B.A. Young (eds.), Proc. 5th Int. Workshop Low Temperature Detectors LTD5, Berkeley, CA, J. Low Temp. Phys. **93**(3/4) (1993) 185–858.
21. H.R. Ott, A. Zehnder (eds.), Proc. 6th Int. Workshop Low Temperature Detectors LTD6, Beatenberg/Interlaken, Switzerland, Nucl. Instrum. Meth. A **370** (1996) 1–284.
22. S. Cooper (ed.), Proc. 7th Int. Workshop Low Temperature Detectors LTD7, Max Planck Institut of Physics Munich, Germany, ISBN 3-00-002266-X, (1997).
23. P. deKorte, T. Peacock (eds.), Proc. 8th Int. Workshop Low Temperature Detectors LTD8, Dalfsen, Netherlands, Nucl. Instrum. Meth. A **444** (2000).
24. F. Scott Porter, D. MacCammon, M. Galeazzi, C. Stahle (eds.). Proc. 9th Int. Workshop Low Temperature Detectors LTD9, American Institute of Physics AIP Conf. Proc. **605** (2002).
25. F. Gatti (ed.), Proc. 10th Int. Workshop Low Temperature Detectors LTD10, Genova, Italy, Nucl. Instrum. Meth. A **520** (2004).
26. M. Ohkubo (ed.), Proc. 11th Int. Workshop Low Temperature Detectors LTD11, Tokyo, Japan, Nucl. Instrum. Meth. A **559** (2006).
27. M. Chapellier, G. Chardin (eds.), Proc. 12th Int. Workshop Low Temperature Detectors LTD12, Paris, France, J. Low Temp. Phys. **151**(1/2, 3/4) (2008).
28. B. Cabrerra, A. Miller, B.Young (eds.), Proc. 13th Int.Workshop Low Temperature Detectors LTD13, Stanford/SLAC, USA, American Inst. of Physics (March 2010).
29. Ch. Enss, A. Fleischmann, L. Gastaldo (eds.), Proc. 14th Int.Workshop Low Temperature Detectors LTD14, Heidelberg, Germany, J.Low Temp.Phys.**167**(5/6) (2012) 561–1196.
30. E. Skirokoff(ed.), Proc. 15th Int.Workshop Low Temperature Detectors LTD15, Pasadena Cal., USA, J. Low Temp. Phys.**176**(3/6)(2014) 131–1108.
31. Ph. Camus, A. Juillard, A. Monfardini (eds.), Proc. 16th Int.Workshop Low Temperature Detectors LTD16, Grenoble, France, J. Low Temp. Phys. **184**(1/4)(2016) 1–978.

32. A. Barone (ed.), Proc. Superconductive Particle Detectors, Torino, Oct. 26–29, 1987, World Scientific.
33. A. Barone, Nucl. Phys. B (Proc. Suppl.) **44** (1995) 645.
34. D. Twerenbold, Rep. Prog. Phy. **59** (1996) 349.
35. H. Kraus, Supercond. Sci. Technol. **9** (1996) 827.
36. N. Booth, B. Cabrera, E. Fiorini, Annu. Rev. Nucl. Part. Sci. **46** (1996) 471.
37. K. Pretzl, *Cryogenic calorimeters in astro and particle physics*, Nucl. Instrum. Meth. A **454** (2000) 114.
38. Ch. Enss (ed.), *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Springer Berlin, Heidelberg, New York (2005).
39. D. McCammon, *Thermal Equilibrium Calorimeters-An Introduction*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 1.
40. J.C. Mather, Appl. Opt. **21** (1982) 1125.
41. S.H. Moseley, J.C. Mather, D. McCammon, J. Appl. Phys. **56**(5) (1984) 1257–1262.
42. D.A. Wollmann et al., Nucl. Instrum. Meth. A **444** (2000) 145.
43. T.C.P. Chui et al., Phys. Rev. Lett. **69**(21) (1992) 3005.
44. B.I. Shklovskii, A.L. Efros, *Electronic Properties of Doped Semiconductors*, Springer-Verlag (1984).
45. P. Colling et al., Nucl. Instrum. Meth. A **354** (1995) 408.
46. K.D. Irwin, Appl. Phys. Lett. **66** (1995) 1998.
47. K.D. Irwin, G.C. Hilton, *Transition Edge Sensors*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 63.
48. B. Young et al., IEEE Trans. Magnetics **25** (1989) 1347.
49. K.D. Irwin, B. Cabrera, B. Tigner, S. Sethuraman, in: Proc. 4th Int. Workshop Low Temperature Detectors for Neutrinos and Dark Matter LTD4, N.E. Booth, G.L. Salmon (eds.), Gif-sur-Yvette, France, Ed. Frontieres (1992), p. 290.
50. P. Ferger et al., Nucl. Instrum. Meth. A **370** (1996) 157.
51. P. Colling et al., Nucl. Instrum. Meth. A **354** (1995) 408.
52. U. Nagel et al., J. Appl. Phys. **76** (1994) 4262.
53. J. Hohne et al., X-Ray Spectrom. **28** (1999) 396.
54. J. Martinis, G. Hilton, K. Irwin, D. Wollmann, Nucl. Instrum. Meth. A **444** (2000) 23.
55. G. Brammertz et al., Appl. Phys. Lett. **80** (2002) 2955.
56. C. Hunt et al., Proc. SPIE **4855** (2003) 318.
57. B. Young et al., Nucl. Instrum. Meth. A **520** (2004) 307.
58. M. Buehler, E. Umlauf, Europhys. Lett. **5** (1988) 297.
59. E. Umlauf, M.Buehler, in: Proc. Int. Workshop Low Temperature Detectors for Neutrinos and Dark Matter LTD4, N.E. Booth, G.L. Salmon (eds.), Gif-sur-Yvette, France, Ed. Frontieres (1992), p. 229.
60. S.R. Bandler et al., J. Low Temp. Phys. **93** (1993) 709.
61. A. Fleischmann et al., Nucl. Instrum. Meth. A **520** (2004) 27.
62. A. Fleischmann, C. Enss, G.M. Seidel, *Metallic Magnetic Calorimeters*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 196.
63. A. Barone, G. Paterno, Physics and Applications of the Josephson Effect, New York, Wiley-Interscience (1984).
64. D. Twerenbold, A. Zehnder, J. Appl. Phys. **61** (1987) 1.
65. H. Kraus et al., Phys. Lett. B **231** (1989) 195.
66. C.A. Mears, S. Labov, A.T. Barfknecht, Appl. Phys. Lett. **63** (21) (1993) 2961.
67. P. Lerch, A. Zehnder, *Quantum Giaever Detectors: STJ's*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 217.
68. N.E. Booth, Appl. Phys. Lett. **50** (1987) 293.

69. K.E. Gray, Appl. Phys. Lett. **32** (1978) 392.
70. I. Giaever, K. Megerle, Phys. Rev. **122**(4) (1961) 1101.
71. S.B. Kaplan et al., Phys. Rev. B **14**(11) (1976) 4854.
72. P.A.J. de Korte et al., Proc. SPIE **1743** (1992) 24.
73. G. Angloher et al., J. Appl. Phys. **89**(2) (2001) 1425.
74. P. Verhoeve et al., Proc. SPIE **6276** (2007) 41.
75. N. Rando et al., Rev. Sci. Instrum. **71**(12) (2000) 4582.
76. R. Christiano et al., Appl. Phys. Lett. **74** (1999) 3389.
77. M.P. Lissitski et al., Nucl. Instrum. Meth. A **520** (2004) 240.
78. E. Figueroa-Feliciano, Nucl. Instrum. Meth. A **520** (2004) 496.
79. P.K. Day et al., Nature **425** (2003) 871.
80. B. Mazin et al., Publ.Astro.Soc. of the pacific **125** (2013) 1348–1361.
81. K. Borer et al., Astroparticle Phys. **22** (2004) 199.
82. K. Borer, M. Furlan, Nucl. Instrum. Meth. A **365** (1995) 491.
83. A. Kotlicki et al., in: Low-Temperature Detectors for Neutrinos and Dark Matter LTD1, K. Pretzl, N. Schmitz, L. Stodolsky (eds.) Springer-Verlag (1987), p. 37.
84. R. Leoni et al., J. Low Temp. Phys. **93**(3/4) (1993) 503.
85. B. van den Brandt et al., Nucl. Phys. B (Proc. Suppl.) **70** (1999) 101.
86. M. Abplanalp et al., Nucl. Instrum. Meth. A **360** (1995) 616.
87. S. Janos et al., Nucl. Instrum. Meth. A **547** (2005) 359.
88. G. Meagher et al., J. Low Temp. Phys. **93**(3/4) (1993) 461.
89. C. Berger et al., J. Low Temp. Phys. **93**(3/4) (1993) 509.
90. S. Calatroni et al., Nucl. Instrum. Meth. A **444** (2000) 285.
91. S. Casalboni et al., Nucl. Instrum. Meth. A **459** (2001) 469.
92. S. Calatroni et al., Nucl. Instrum. Meth. A **559** (2006) 510.
93. M. Abplanalp, Nucl. Instrum. Meth. A **370** (1996) 11.
94. K. Pretzl, Particle World **1**(6) (1990) 153.
95. K. Pretzl, J. Low Temp. Phys. **93** (1993) 439.
96. F. Zwicky, Helv. Phys. Acta **6** (1933) 110.
97. S. Perlmutter et al., Astrophys. J. **483** (1997) 565.
98. A.G. Riess et al., Astronom. J. **116** (1998) 1009.
99. The Planck collaboration (P.Ade et al.), Astronom. Astrophys. **594**A (13) (2016) and arXiv: 1502.01589 (astro-phys.Co) (2016)
100. R.D. Peccei, H.R. Quinn, Phys. Rev. Lett. **38** (1977) 1440.
101. P.W. Graham et al., An.Rev.Nucl. and Particle Searches **65** (2015) 485–514 and arXiv: 1602.00039 (hep - ex) (2016)
102. F.Kahlhofer, Int.J.Mod.Phys. A **32** (2017) 1730006 and arXiv: 1702.02430 (hep - ph)
103. K. Pretzl, Space Science Reviews **100** (2002) 209.
104. G. Jungman, M. Kamionkowski, K. Griest, Phys. Rep. **267** (1996) 195.
105. B. Sadoulet, in: Low Temperature Detectors for Neutrinos and Dark Matter LTD1, K. Pretzl, L. Stodolsky, N. Schmitz (eds.), Springer-Verlag (1987), p. 86.
106. L. Gonzalez-Mestres, D. Perret-Gallix, Nucl. Instrum. Meth. A **279** (1989) 382.
107. S. Cebrian et al., Phys. Lett. B **563** (2003) 48.
108. P. Meunier et al., Appl. Phys. Lett. **75**(9) (1999) 1335.
109. D.S. Akerib et al., Phys. Rev. D **72** (2005) 052009.
110. Z.Ahmed et al., Phys. Rev. Lett. **106** (2011) 131302 and R.Agnese et al., Phys. Rev. D **92** (7) (2015) 072003 and arXiv:1504.05871 (hep-ex).
111. R.Agnese et al., Phys. Rev. Lett. **111** (2013) 251301 and arXiv:1304.4279 (hep-ex).
112. N.Luke, J. Appl. Phys. **64** (1988) 6858.
113. G. Wang, J. Appl. Phys. **107** (2010) 094504.
114. C. Isaila et al.,Phys. Lett. B **716** (2012) 160.
115. R. Agnese et al.,Phys. Rev. Lett. **116** (7) (2016) 071301, R.Agnese et al., arXiv:1707.01632 (2017) submitted to Phys. Rev. D.
116. V. Sanglard et al., Phys. Rev. D **71** (2005) 122002.

117. E.Armengaud et al.,Phys.Lett. B **702** (2011) 329–335 and arXiv:1103.4070 (astro-ph.Co) (2011)
118. L.Hehn et al., Eur. Phys. J. C **76** (10) (2016) 548 and arXiv: 1607.03367 (astro-ph.Co) (2016), E.Armengaud et al., arXiv 1706.01070 (physics.ins-det) (2017)
119. T. Shutt et al., Nucl. Instrum. Meth. A **444** (2000) 34.
120. D.S. Akerib et al., Phys. Rev. D **68** (2003) 082002.
121. G. Angloher et al., Astroparticle Physics **31** (2009) 270–276 and arXiv:0809.1829 [astro-ph].
122. C. Isaila et al., Nucl. Instrum. Meth. **559** (2006) 399; and C. Isaila et al., J. Low Temp. Phys. **151**(1/2) (2008) 394.
123. J. Ninković et al., Nucl. Instrum. Meth. A **564** (2006) 567.
124. G. Angloher et al., Eur.Phys. J.C **76** (1) (2016) 25 and arXiv: 1509.01515.
125. R. Strauss et al., Eur.Phys. J.C bf77 (2017) 506 and arXiv: 1704.04320 (physics.ins-det) (2017).
126. R. Strauss et al., Phys.Rev. D bf96 (2) (2017) 022009 and arXiv: 1704.04317.
127. D. Akimov et al., Science **357** (2017) 1123
128. Mark Schumann, Physics Departement University Freiburg Germany, private communication.
129. A. Aguilar-Arevalo et al., Phys. Rev. D **94** (2016) 082006 and arXiv: 1607.07410.
130. E. Aprile et al., arXiv: 1705.06655.
131. D. Akerib et al., Phys. Rev. Lett. **118** (2) (2017) 021303 and arXiv: 1608.07648.
132. A. Tan et al., Phys. Rev. Lett. **117** (12) (2016) 12133 and arXiv: 1708.06917.
133. E. Aprile et al., Phys. Rev. D **94** (12) (2016) 122001 and arXiv: 1609.06154.
134. P. Agnes et al., Phys. Lett. B **743** (2015) 456–466 and arXiv:1410.0653.
135. R.Bernabei et al., Eur. Phys.J.Web Conf **13** (2017) 60500 and arXiv:1612.01387 R.Bernabei et al. Eur. Phys. J. C **73** (2013) 2648.
136. C. E. Aalseth et al., Phys. Rev. Lett. **106** (13) (2011) 131301 and arXiv:1401.3295.
137. F.Froborg et al.,arXiv:1601.05307.
138. E. Behnke et al., Astropart. Phys. **90** (2017) 85–92.
139. G. Agloher et al.,Phys.Dark Univ. **3** (2014) 41–74.
140. R.Agnese et al.,Phys. Rev. D **95** (8) (2017) 082002 and arXiv: 1610.0006.
141. J.Aalbers et al., JCAP **11** (2016) 017 and arXiv: 1606.07001.
142. D.S.Akerib et al., Astropart. Phys. **96** (2017) 1–10
143. Y. Fukuda et al., Phys. Rev. Lett. **81**(8) (1998) 1562.
144. M. Maltroni, T. Schwetz, M. Tortola, J.W.F. Valle, New J. Phys. **6** (2004) 122.
145. M. Goeppert Mayer, Phys. Rev. **48** (1935) 512.
146. P. Vogel, *Double Beta decay: Theory, Experiment and Implications*, in: *Current aspects of Neutrino Physics*, D.O. Caldwell (ed.), Springer-Verlag (2001), p. 177.
147. C. Arnaboldi et al., Phys. Rev. Lett. **95** (2005) 142501.
148. K.Alfonso et al., Phys. Rev. Lett. **115** (10) (2015) 102502 and arXiv: 1504.02454 (nucl.-ex).
149. C.Alduino et al., J.INST **11** (7) (2016) P07009 and arXiv: 1604.05465 (phys.ins-det) (2016).
150. C.Alduino et al., Eur. Phys. J. C **77** (8) (2017) 532 and arXiv: 1705.10816 (phys.ins-det) (2017).
151. G.Wang et al., arXiv: 1504.03599 (phys. ins.-det) (2015)
152. A.Gaudo et al., Phys. Rev. Lett. **110** (2013) 062502 and arXiv: 1605.02889 (2016)
153. J.B. Albert et al., Nature **510** (2014) 229
154. M. Agostini et al., Nature **544** (2017) 5
155. R.Arnold et al., Phys. Rev. D **92** (2015) 072011
156. S.Betts et al., arXiv: 1307.4738 (astro-ph.IM) (2013)
157. V.M. Lobashov, Nucl. Phys. A **719** (2003) 153, and references therein.
158. KATRIN experiment, https//www.katrin.kit.edu.
159. F. Fontanelli, F. Gatti, A. Swift, S. Vitale, Nucl. Instrum. Meth. A **370** (1996) 247.
160. F. Gatti et al., Nucl. Phys. B **91** (2001) 293.

161. M. Sisti et al., Nucl. Instrum. Meth. A **520** (2004) 125.
162. F. Gatti et al., Nature **397** (1999) 137.
163. F. Gatti, F. Fontanelli, M. Galeazzi, S. Vitale, Nucl. Instrum. Meth. A **444** (2000) 88.
164. C. Arnaboldi et al., Phys. Rev. Lett. **96** (2006) 042503.
165. A. Nucciotti, J. Low Temp. Phys. **151**(3/4) (2008) 597.
166. E. Ferri et al.,J. Low Temp. Phys. **176** (5/6) (2014) 885–890.
167. L. Gastaldo et al.,J. Low Temp. Phys. **176** (5/6) (2014) 876–884.
168. B.Alperit et al., The European Physical Journal C **75** (2015) 112.
169. M.P. Croce et al., J.Low Temp. Phys. **184** (3/4) (2016) 958–968 and arXiv:1510.03874.
170. A. de Rujula and M. Lusignoli, Phys. Lett. B **118** (1982) 429434.
171. O. Dragoun and D. Vnos, Open Physics Journal **3** (2016) 73–113.
172. D. McCammon et al., Astrophys. J **576** (2002) 188.
173. C.K. Stahle et al., Nucl. Instrum. Meth. A **520** (2004) 466.
174. D.D.E. Martin et al., Nucl. Instrum. Meth. A **520** (2004) 512.
175. N. Rando et al., J. Low Temp. Phys. **93**(3/4) (1993) 659.
176. B. Cabrera et al., Appl. Phys. Lett. **73** (1998) 735.
177. D.D.E. Martin et al., Proc. SPIE **6269** (2006) 62690O-1.
178. J. Burney et al., Nucl. Instrum. Meth. A **559** (2006) 506.
179. R.W. Romani et al., Astrophys. J. **563** (2001) 221.
180. J.A. Chervanek et al., Appl. Phys. Lett. **44** (1999) 4043.
181. P.A.J. de Korte et al., Rev. Sci. Instrum. **74**(8) (2003) 3807.
182. W.B. Doriese et al., Nucl. Instrum. Meth. A **559** (2006) 808.
183. R.A. Hijemering et al., Nucl. Instrum. Meth. A **559** (2006) 689.
184. B. Cabrera, R. Romani, *Optical/UV Astrophysics Applications of Cryogenic detectors*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 416.
185. P. Verhoeve, J. Low Temp. Phys. **151**(3/4) (2008) 675.
186. M.J.Strader et al. Astrophys. J. Letters **779** (2013) L12.
187. Asad M. Aboobaker et al., arXiv: 1703.03847 (astro-phys.IM) (2017).
188. J.M.Nagy et al., Astrophysics J. **844** (2) (2017) 151 and arXiv:1704.0025.
189. P.Ade et al., The Astrophysical Journal **794** (171) (2014) 21.
190. P.A.R. Ade et al., Phys. Rev. Lett. **112** (24) (2014) 241101.
191. PLANCK Collaboration, Astronomy & Astrophysics **586** (2014) A133.
192. P.A.R. Ade et al. Phys. Rev. Lett. **114** (10) (2015) 101301 and arXiv:1502.00612.
193. L. Lesyna et al., J. Low Temp. Phys. **93** (1993) 779.
194. R. Ladbury, Physics Today (July 1998) 19.
195. D.E. Newbury et al., *Electron Probe Microanalysis with Cryogenic Detectors*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 267.
196. V. Shvarts et al., Nucl. Instrum. Meth. A **520** (2004) 631.
197. D. Twerenbold, Nucl. Instrum. Meth. A **370** (1996) 253.
198. M. Frank et al., Mass Spectrometry Reviews **18** (1999) 155.
199. D. Twerenbold et al., Proteomics **1** (2001) 66.
200. M. Frank et al., Rapid Commun. Mass Spectrom. **10**(15) (1996) 1946.
201. G.C. Hilton et al., Nature **391** (1998) 672.
202. J.N. Ullom, J. Low Temp. Phys. **151**(3/4) (2008) 746.
203. S. Rutzinger et al., Nucl. Instrum. Meth. A **520** (2004) 625.
204. P. Christ et al., Eur. Mass Spectrom. **10** (2004) 469.
205. P. Egelhof, S. Kraft-Bermuth, *Heavy Ion Physics*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 469.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 20
Detectors in Medicine and Biology

P. Lecoq

20.1 Dosimetry and Medical Imaging

The invention by Crookes at the end of the nineteenth century of a device called spinthariscope, which made use of the scintillating properties of Lead Sulfide allowed Rutherford to count α particles in an experiment, opening the way towards modern dosimetry. When at the same time Wilhelm C. Roentgen, also using a similar device, was able to record the first X-ray picture of his wife's hand 2 weeks only after the X-ray discovery, he initiated the first and fastest technology transfer between particle physics and medical imaging and the beginning of a long and common history.

Since that time, physics, and particularly particle physics has contributed to a significant amount to the development of instrumentation for research, diagnosis and therapy in the biomedical area. This has been a direct consequence, one century ago, of the recognition of the role of ionizing radiation for medical imaging as well as for therapy.

20.1.1 Radiotherapy and Dosimetry

The curative role of ionizing radiation for the treatment of skin cancers has been exploited in the beginning of the twentieth century through the pioneering work of some physicists and medical doctors in France and in Sweden. This activity has very much progressed with the spectacular developments in the field of accelerators,

P. Lecoq (✉)
CERN, Geneva, Switzerland
e-mail: Paul.Lecoq@cern.ch

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_20

beam control and radioisotope production. Today radiotherapy is an essential modality in the overall treatment of cancer for about 40% of all patients treated. Conventional radiotherapy (RT) with X-rays and electrons is used to treat around 20,000 patients per 10 million inhabitants each year.

The main aim of radiation therapy is to deliver a maximally effective dose of radiation to a designated tumour site while sparing the surrounding healthy tissues as much as possible. The most common approach, also called teletherapy, consists in bombarding the tumour tissue with ionizing radiation from the outside of the patient's body. Depending on the depth of the tumour, soft or hard X-rays or more penetrating γ-rays produced by a ^{60}Co source or by a linac electron accelerator are used. However, conventional X-ray or γ-ray radiation therapy is characterized by almost exponential attenuation and absorption, and consequently delivers the maximum energy near the beam entrance. It also continues to deposit significant energy at distances beyond the cancer target. To compensate for the disadvantageous depth-dose characteristics of X-rays and γ-rays and to better conform the radiation dose distribution to the shape of the tumour, the radiation oncologists use complex Conformal and Intensity Modulated techniques (IMRT) [1]. The patient is irradiated from different angles, the intensity of the source and the aperture of the collimators being optimized by a computer controlled irradiation plan in order to shape the tumour radiation field as precisely as possible.

Another way to spare as much as possible healthy tissues is to use short range ionizing radiation such as β or α particles produced by the decay of unstable isotopes directly injected into the tumour. This method, called brachytherapy, has been originally developed for the thyroid cancer with the injection of ^{153}I directly into the nodules of the thyroid gland. It is also used in other small organs such as prostate or saliva gland cancer.

Following these trends a new generation of minimally invasive surgical tools appears in hospitals, allowing to precisely access deep tumours from the exterior of the body (gamma-knife), or by using brachytherapy techniques, i.e. by injecting radioisotopes directly in the tumour (beta-knife, alpha-knife and perhaps soon Auger-knife). More recently the radio-immunotherapy method has been successfully developed: instead of being directly inserted in the tumour, the radioactive isotopes can be attached by bioengineering techniques to a selective vector, which will bind to specific antibody receptors on the membranes of the cells to be destroyed. Typical examples are the use of the α emitter ^{213}Bi for the treatment of leukaemia and of the β-emitter ^{90}Y for the treatment of glioblastoma.

In 1946 Robert Wilson, physicist and founder of Fermilab, proposed the use of hadron beams for cancer treatment. This idea was first applied at the Lawrence Berkeley Laboratory (LBL) where 30 patients were treated with protons between 1954–1957. Hadrontherapy is now a field in rapid progress with a number of ambitious projects in Europe, Japan and USA [2], exploiting the attractive property of protons and even more of light ions like carbon to release the major part of their kinetic energy in the so-called Bragg peak at the end of their range in matter (Fig. 20.1).

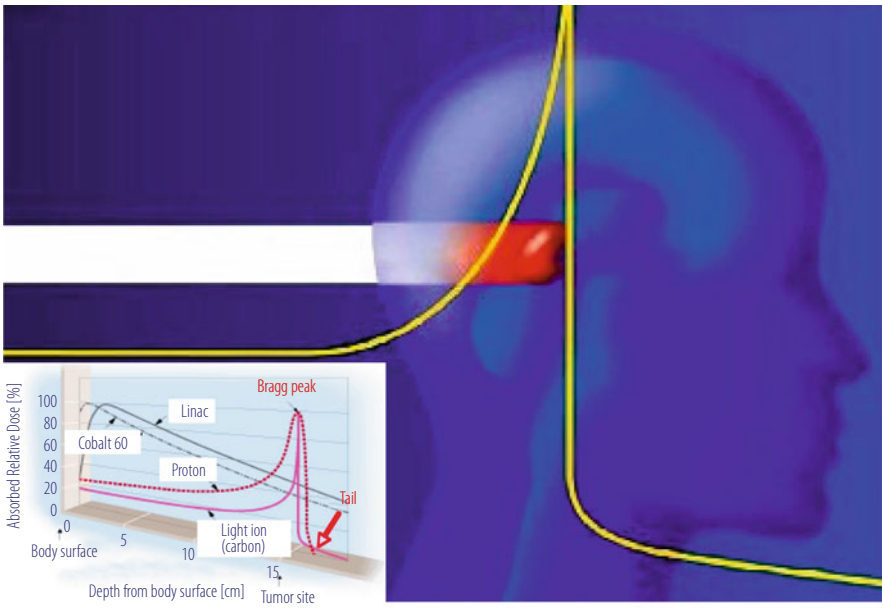


Fig. 20.1 Bragg peak for an ion beam in the brain of a patient. The insert shows the energy absorbed by tissues as a function of depth for different radiation sources (Courtesy U. Amaldi)

It is particularly important to treat the disease with the minimum harm for surrounding healthy tissues. In the last centimeters of their range the Linear Energy Transfer (LET) of protons or even more of carbon ions is much larger than the one of X-rays (low-LET radiations). The resulting DNA damages include more complex double strand breaks and lethal chromosomal aberrations, which cannot be repaired by the normal cellular mechanisms. The effects produced at the end of the range are therefore qualitatively different from those produced by X- or γ-rays and open the way to a strategy to overcome radio-resistance, often due to hypoxia of the tumour cells. For these reasons carbon ions with their higher relative biological effectiveness (RBE) at the end of their range, of around a factor of three higher than X-rays, can treat tumours that are normally resistant to X-rays and possibly protons. This treatment is particularly applicable to deep tumours in the brain or in the neck as well as ocular melanoma.

Whatever the detailed modality of the treatment planning, precise dosimetry is mandatory to develop an optimal arrangement of radiation portals to spare normal and radiosensitive tissues while applying a prescribed dose to the targeted disease volume. This involves the use of computerized treatment plan optimization tools achieving a better dose conformity and minimizing the total energy deposition to the normal tissues (Fig. 20.2). It requires a precise determination and simulation of the attenuation coefficients in the different tissues along the beam. These data are obtained from high performance anatomic imaging modalities such as X-ray computed tomography (CT) and magnetic resonance (MRI).

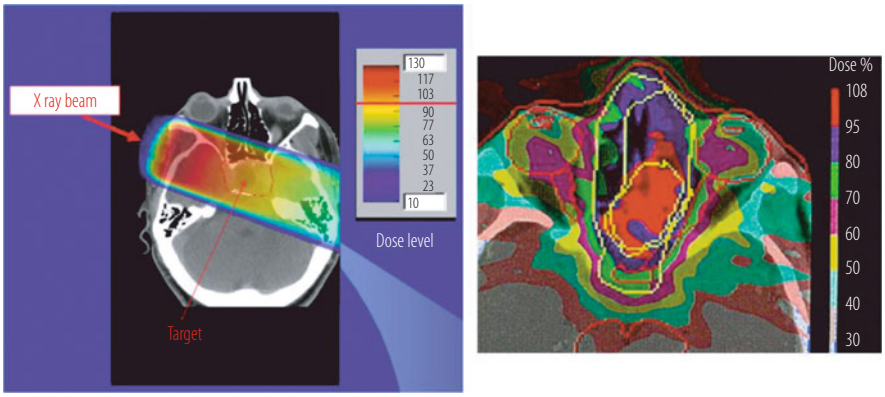


Fig. 20.2 Dosimetry for a brain tumour in the case of one (left) or nine crossed (right) X-ray beams. The treatment plan is based on a tumour irradiation of at least 90 Gy (Courtesy U. Amaldi)

For the particular case of hadrontherapy on-line dosimetry in the tissues is in principle possible. It relies on the production of positron emitter isotopes produced by beam spallation (^{10}C and ^{11}C for 12C beam) or target fragmentation during the irradiation treatment. The two 511 keV γ produced by the positron annihilation can be detected by an in-line positron emission tomography (PET) to precisely and quantitatively map the absorbed dose in the tumour and surrounding tissues. Although challenging because of the timing and high sensitivity requirements this approach is very promising and a number of groups are working on it worldwide [3].

20.1.2 Status of Medical Imaging

The field of medical imaging is in rapid evolution and is based on five different modalities: X-ray radiology (standard, digital and CT), isotopic imaging (positron emission tomography, PET, and single photon emission computed tomography SPECT), ultrasound (absorption, Doppler), magnetic resonance (MRI, spectroscopy, functional), and electrophysiology with electro- and magneto-encephalography (EEG and MEG). More recently, direct optical techniques like bioluminescence and infrared transmission are also emerging as powerful imaging tools for non-too-deep organs.

For a long time imaging has been anatomical and restricted to the visualization of the structure and morphology of tissues allowing the determination of morphometric parameters. With the advent of nuclear imaging modalities (PET and SPECT) and of the blood oxygen level dependent (BOLD) technique in magnetic resonance imaging (MRI) functional imaging became possible and medical doctors can now see organs at work. Functional parameters are now accessible in vivo and in real

time, such as vascular permeability, haemodynamics, tissue oxygenation or hypoxia, central nervous system activity, metabolites activity, just to cite a few.

In the current clinical practice medical imaging is aiming at the in-vivo anatomic and functional visualization of organs in a non- or minimally invasive way. Isotopic imaging, in particular PET, currently enjoys a spectacular development. Isotopic imaging consists in injecting into a patient a molecule involved in a specific metabolic function so that this molecule will preferentially be fixed on the organs or tumours where the function is at work. The molecule has been labeled beforehand with a radioisotope emitting gamma photons (Single Photon Emission Computed Tomography or SPECT) or with a positron emitting isotope (Positron Emission Tomography or PET). In the latter case, the positron annihilates very quickly on contact with ordinary matter, emitting two gamma photons located on the same axis called the line of response (LOR) but in opposite directions with a precise energy of 511 keV each. Analyzing enough of these gamma photons, either single for SPECT or in pairs for PET, makes it possible to reconstruct the image of the area (organ, tumour) where the tracer focused.

Since the beginning to the twenty-first century a new generation of machines became available, which combine anatomic and functional features: the PET/CT. This dual modality system allows the superposition of the high sensitivity functional image from the PET on the precise anatomic picture of the CT scanner. PET/CT has now become a standard in the majority of hospitals, particularly for oncology. This trends for multimodal imaging systems is increasing both for clinical and for research applications (Fig. 20.3).

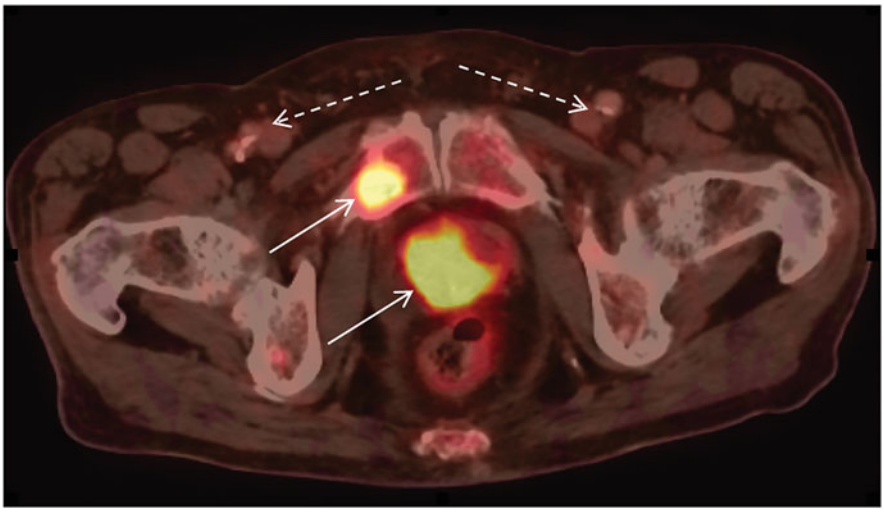


Fig. 20.3 Abdominal slice of a 78 year-old male, with biopsy-proven prostate adenocarcinoma and penile adenocarcinoma. Focal uptake in the prostate bed and in the penile shaft (full arrows). Multiple foci in the pelvis compatible with skeletal metastases (dashed arrows) (Courtesy D. Townsend)

Table 20.1 Comparison of the performances of four imaging modalities

Imaging modality	Type of imaging	Examination time	Spatial resolution
PET	Functional and molecular (picomolar sensitivity)	10–20′ (whole body)	3–5mm
SPECT	Functional and molecular	10–20′ (per 40 cm filed)	6–8 mm
MRI	Anatomical Functional (millimolar sensitivity)	30–60′	0.5 mm
CT	Anatomical	<1′ (whole body)	0.5 mm

The most frequently used positron emitters are ^{18}F, ^{11}C, ^{15}O, ^{13}N, the three last ones being isotopes of the nuclei of organic molecules.

As compared to other non-invasive imaging modalities isotopic imaging has a functional sensitivity at the picomolar level, which is several orders of magnitude better than magnetic resonance. It opens incredible perspectives for cell and molecular imaging, in particular for visualizing and quantifying genomic expression or tissue repair efficiency of stem cells. However, the detection efficiency compared to the dose injected to the patient, also called "sensitivity", is strongly limited by technical constraints, and the spatial resolution is still one order of magnitude worse than for CT or MRI (Table 20.1).

20.1.3 Towards In-Vivo Molecular Imaging

The challenge of future healthcare will be to capture enough information from each individual person to prevent disease at its earliest stage, to delineate disease parameters, such as aggressiveness or metastatic potential, to optimize the delivery of therapy based on the patient's current biologic system and to quickly evaluate the treatment therapeutic effectiveness.

New therapeutic strategies are entering the world of major diseases. They aim to acquire as fast as possible all the information on the pathological status of the patient in order to start adapted therapeutics and therefore to minimize the handicap. This applies to neurological and psychiatric diseases but also to the treatment of inflammatory diseases, such as rheumatologic inflammation, and of cancer. Moreover, the non-invasive determination of the molecular signature of cancers in the early stage of their development, or even before the tumour growth, will help to target the therapeutic strategy and to reduce considerably the number of unnecessary biopsies.

This trend is supported by the new paradigm of "personalized medicine" (also called precision medicine), which aims at delivering "the right treatment, to the right patient, at the right time". Personalised medicine refers to a medical model using

characterisation of individuals' phenotypes and genotypes (e.g. molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention. In this new healthcare context, a radical shift is currently taking place in the way diseases are managed: from the present one-fits-all approach to one that delivers medical care tailored to the needs of individual patients. This includes the detection of disease predisposition, early diagnosis, prognosis assessment, measurement of drug efficacy and disease monitoring. To achieve this ambitious goal, there is an increased demand for simultaneous in-vivo quantitative and dynamic characterization of several biological processes at the molecular and genetic level. A new generation of whole body and organ-specific imaging devices is needed combining the excellent sensitivity and specificity of PET or SPECT with a high-spatial resolution imaging modality (CT, MR optical or US) providing additional functional, metabolic or molecular information.

For many years, physicians relied on the use of anatomical imaging to non-invasively detect tumours and follow up their growth. Functional imaging such as bone or thyroid scintigraphy and more recently PET using 18FDG for example, has provided more information for tumour staging. The next revolution being prepared will have to do with molecular imaging. The goal is in-vivo visual representation, characterization and quantification of biological processes at the cellular and sub-cellular level within living organisms. This is the challenge of modern biology: detect early transformations in a cell, which may lead to pathology (precancerous activity, modifications of neuronal activity as warning signs of Alzheimer or Parkinson disease). Besides early detection, assessment of prognosis and potential response to therapy will allow a better treatment selection through a precise delineation of molecular pathways from genes to disease. All aspects of gene expression will be addressed (genomics, proteomics, transcriptomics, enzymatic activity), but also the molecular signal transduction through cell membranes (a key to determine the efficacy of drugs) as well as the identification and quantification of specific cell receptors over-expressed in some pathological situations, such as dopamine receptors for epilepsy.

With the development of new imaging probes and "smart probes", imaging provides cellular protein and signal-pathway identification. There is an increasing amount of molecular probes dedicated to imaging but also to tumour therapy. The molecular phenotype of cells composing the tumour can lead to tailored therapies. This tumour phenotype can be determined ex-vivo on tissue samples. Molecular imaging should allow performance of an in-vivo tumour phenotyping by an appropriate use of specific imaging probes. This molecular profiling could already be envisioned in the very near future for some specific tumours overexpressing peptide hormone receptors such as breast and prostate cancers, and should become widely developed.

Therefore, it represents a major breakthrough to provide the medical community with an integrated "one-stop-shop" molecular profiling imaging device, which could detect tracers dedicated to Single Photon Emission Computed Tomography

(SPECT) or PET, as well as Magnetic Resonance Imaging (MRI), or X-ray Computed Tomography (CT) contrast agents.

Furthermore, since functional imaging allows the assessment of biochemical pathways, it will also provide accurate tools for experimental research. As an example, a large effort worldwide has recently allowed the precise mapping of the different genes in the DNA sequence but the mechanisms, by which these genes produce proteins, interact with each other, regulate their expression, are far from understood. In other terms we can say that the genomic alphabet has been decoded but its dynamic expression, its grammar, remains to be studied and understood. In-vivo molecular imaging of gene expression is now within reach through the development of ever more elaborated molecular probes as well as of sophisticated techniques which significantly improve the performances of modern imaging devices.

Drug development also takes advantage of technical progress in imaging technologies, like quantitative positron emission tomography in small animals, to determine drug pharmacokinetics and whole body targeting to tissues of interest. Moreover, the combination of functional imaging with a high resolution anatomical method such as MRI and/or X-ray CT will considerably enhance the possibility of determining the long term efficiency of a drug on basic pathological processes such as inflammation, blood flow, etc. In particular, the expected progress in sensitivity, timing and spatial resolution, coupled with a true multi-modality registration, will allow to explore the activity of a drug candidate or other essential pathophysiological processes of disease models in animals, like for instance cancer or adverse inflammatory effects.

This approach will require targeting cellular activity with specific contrast agents, but also a large effort on imaging instrumentation. Developments are needed for faster exams, correction of physiological movements during acquisition time (breathing, cardiac beating, digestive bolus), access to dynamic processes, quantification, true multimodality, dose reduction to the patient. This will require significant improvements in spatial and timing resolution, sensitivity and signal-to-noise ratio, all parameters very familiar to particle physicists. From the technologies already available, developed for instance for the LHC detectors or under development for the future linear collider, fast crystals, highly integrated fast and low noise electronics and ultrafast Geiger mode SiPMs open the way to time of flight (TOF) PET. These technologies are progressively being implemented in commercial PET's, resulting in an improvement of image signal/noise ratio with a corresponding sensitivity increase. Sensitivity to picomolar concentrations are within reach for whole body commercial PET scanners, which correspond to the molecular activity of a few hundreds of cells only.

20.2 X-Ray Radiography and Computed Tomography (CT)

20.2.1 Different X-Ray Imaging Modalities

X-ray radiology is the most popular imaging technique, which comprises X-ray Radiography, Computed Tomography (CT), also called Tomo-Densitometry (TDM), and Dual Energy X-ray Absorptiometry (DXA). For planar radiography the general trend is to progressively replace the film by digital devices, as already used for CT. The patient is exposed to an X-ray source, with its energy being adjusted as a function of the density of the tissues to be visualized. Present systems work in signal integration mode, although there is a trend towards photon counting devices, as will be explained in Sect. 20.2.4. In standard radiography, the projected image is recorded either on a photographic plate or on a digital device using a scintillation material coupled to a photosensitive array of Silicon diodes. Computed tomography or Tomo-Densitometry is based on the detection of X-ray attenuation profiles from different irradiation directions. Both the X-ray source and the detector (usually an array of scintillators coupled to a solid-state photodetector is rotating around the patient as the bed is moving through the scanner. This technique allows a 3D reconstruction of attenuation density within the human body. These density profiles can then be viewed from different directions and analyzed in a succession of slices allowing a full 3D reconstruction of the anatomical image (Fig. 20.4).

20.2.2 Detection System

The X-rays to be detected must be first converted into visible light in a scintillator or into electron-hole pairs in a semiconductor device, which are directly recorded. The X-rays are absorbed in a phosphor screen, in which they excite different luminescent centres depending on the nature of the phosphor. The visible light produced by these luminescent centres is recorded on an emulsion deposited on a film or a photographic plate or on a photodiode array in direct contact with the scintillating screen or through an optical relay lens system. For about one century, film has been the unique tool for X-ray radiography. There is a trade-off between the thickness of the phosphor screen, which has to be thick enough to efficiently absorb X-rays, but not too much in order to minimize the light spread and image blurring caused by the distance between the light emission point on the screen and the recording emulsion. To take advantage of the exponential absorption of X-rays in the screen causing a larger number of X-ray absorbed at the entrance of the phosphor screen, the film is generally placed in front of the screen in a so-called back screen configuration. Soft tissues, characterized by low X-ray absorption, are seen as bright areas on the phosphor screen because of the large number of X-rays reaching the screen. The visible light photons are absorbed in the emulsion where they convert (after development of the latent image) the silver halide grains into metallic silver. As a

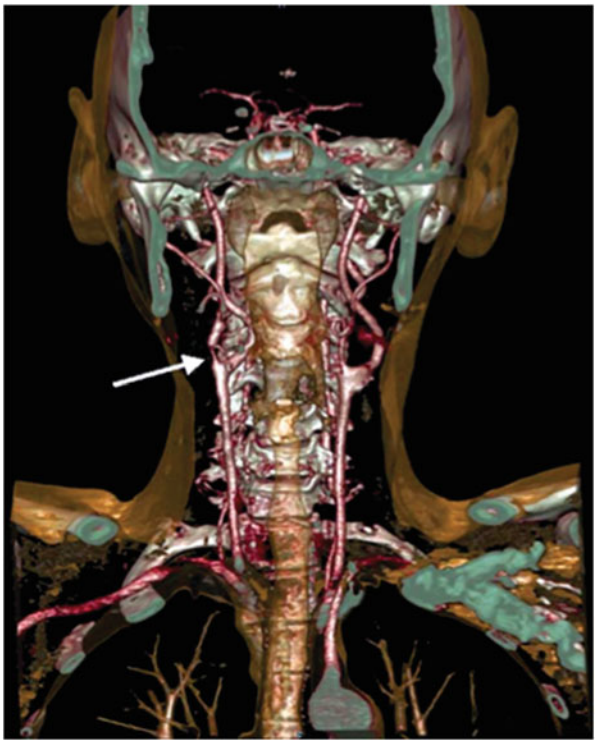


Fig. 20.4 64-Slice CT of the carotid arteries and circle of Willis of a patient. The arrow indicates a severe stenosis (Courtesy of Siemens Medical Solutions)

result, soft tissue produce black images whereas denser parts of the body like bones appear clear.

Digital radiography has progressively been replacing film-based radiography. Indeed, it offers a number of advantages such as better linearity, higher dynamic range, and most importantly, the possibility of distributed archiving systems. Besides direct conversion detectors like amorphous Silicon, CdTe or CdZnTe, which will be described in Sect. 20.2.4, scintillation materials are still the detectors of choice for modern X-ray detectors. For thin scintillation screens (0.1–0.2 mm thickness), which are well adapted to the lowest X-ray energies (for instance about 20 keV for X-ray mammography), ceramic phosphors are commonly used because they can be produced in any shape at a reasonable cost. On the other hand, for dental X-ray diagnostics (about 60 keV) and full body X-ray computed tomography (about 150 keV) the required stopping power would require much thicker screens and monocrystalline inorganic scintillators have been generally preferred up to now because of their much higher light transparency than ceramics. However, recent progress in producing more transparent ceramics based on nanopowders with low dispersion grain diameter may change this situation.

20.2.2.1 Scintillators for X-Ray Conversion

Detector elements of old CT scanners were prevalently implemented as ionization chambers filled with xenon at high pressure. Such detectors usually absorb 30–40% of the impinging photons and generate about 5500–6000 electrons per photon of 100 keV. Modern digital radiography devices and CT scanners use scintillator material arrays optically coupled to matching silicon p-i-n photodiode matrices. The scintillating material must be sufficiently thick to absorb close to 100% of the impinging photons, thus minimizing the patient dose required for a given image quality. Latest generation X-ray CT scanners are recording about 1000 projections (subject slices) per second. This imposes severe constrains on both the decay time and afterglow of the scintillating material. Afterglow is known to produce ghost images through a "memory effect" which deteriorates the quality of the images. The requirements for the scintillator material to be used in X-ray CT are:

- High absorption for X-rays in the energy range up to 140 keV. Absorption close to 100% for ~2 mm thick material layer is required to achieve an acceptable X-ray CT image to noise ratio. Indeed, the image quality is limited in low contrast regions by statistical fluctuations in the numbers of detected X-rays. A high detection efficiency allows to keep the patient dose exposure within reasonable limits for a given image quality. With last generation CT scanners, a whole body CT scan can now be achieved with a dose of less than 5 mSv, close to the level of 1 year natural radioactivity exposure.
- High light output, typically of the order or greater than 20,000 photons/MeV in order to reduce the image noise relative to (low) signal levels.
- Radioluminescence spectrum in the visible or near IR range to match the spectral sensitivity of the silicon photodetectors.
- Decay time in the range of 1–10 μs, in order to achieve sampling rates of the CT scanners in the $\geq$10 kHz range.
- Very low afterglow. Afterglow is generally caused by material imperfections (impurities, defects), causing delayed detrapping and carrier recombination with decay times in the range 100 ms to 10 s. An afterglow level of less than 0.1% is generally required 3 ms after the end of a continuous X-ray excitation. Afterglow causes blurring in the CT images.
- Good radiation hardness. The integrated exposure of the scintillators can reach several tens of kGy over the lifetime of a CT scanner. Changes in the light yield cause detector gain instability, resulting in image artifacts. Long-term changes of ~10% are acceptable, while only less than 0.1% short term changes during the daily operation (1000 R) can be tolerated without image quality degradation.
- Small temperature dependence of the light yield. The X-ray generation system usually dissipates a high amount of energy and the temperature of the detectors can change rapidly. A light output temperature coefficient within ±0.1%/°C is desirable, which is a rather stringent requirement. Cadmium Tungstate (the most frequently used crystal in modern commercial CT scanners) has an acceptable temperature coefficient of −0.3%/°C [4].

- Good mechanical properties allowing micromachining of 2D scintillator arrays with pixel dimensions less than 1 mm.
- Affordable cost.

Table 20.2 summarizes the main characteristics of the scintillators used in medical CT imaging. During the last decade, there was a clear trend towards synthesized ceramic scintillators [5].

The only crystalline material still in use in medical and security systems CT scanners is cadmium tungstate, $CdWO_4$, also called CWO. Its main advantage over CsI(Tl) is a very low afterglow level of 0.05% 3 ms after the end of the X-ray exposure and a reasonable temperature coefficient of 0.3%/°C. In spite of their wide use CWO crystals are however not optimal for CT applications due to their brittleness and the toxicity of cadmium. Moreover, it is difficult to manufacture crystals with adequate uniformity. This has been an argument for the search of a new generation of CT scintillators. This search was initiated by General Electric and Siemens in the mid of the 1980s when they introduced the first polycrystalline ceramic scintillators. The host materials are yttrium and gadolinum oxides: Y_2O_3 and Gd_2O_3, which, after doping with Pr and Tb, demonstrate reasonable scintillation properties. However, their transmission is rather low, ceramics being more translucent than transparent. The additional Eu^{3+} activator efficiently traps electrons to form a transient Eu^{2+} state, allowing holes to form Pr^{4+} and Tb^{4+} and, therefore, competes with the intrinsic traps responsible for afterglow. This energy trapped on the Pr and Tb sites decays non-radiatively in presence of the Eu ions reducing therefore the level of afterglow [6].

Gadolinium oxide ceramic is now replaced by yttrium gadolinum oxide YGO [7], and gadolinum silicate GOS based ceramic materials [8]. When coupled to a silicon p-i-n photodiode they generate about 20 electrons per 1 keV of absorbed X-ray energy. However the long decay time of YGO (~1 ms) is a major concern and requires a complex algorithm of data deconvolution to suppress the effects of afterglow at the price of increased projection noise. Other ceramic materials proposed for CT applications are gadolinium gallium garnet, and lanthanum hafnate [9]. While ceramic materials are generally preferred to crystals because of their good performance and easy production in a variety of shapes, their low transparency requires the use of thin scintillators elements, with lower than optimal X-ray efficiency.

A large R&D effort is under way by several companies to produce flat panels for digital radiography. The standard scintillating crystal or ceramic pixels are replaced by detector arrays made of CsI(Tl) needles or small crystals (e.g. calcium tungstate CWO or YAP) directly coupled to photodiode arrays or segmented photomultipliers (see next section).

Table 20.2 Properties of scintillators used in X-ray CT imaging

Scintillator	Density (g/cm^3)	Thickness to stop 99% of 140 keV X-rays (mm)	Light yield (ph/MeV)/Temperature coefficient (%/°C)	Peak of emission band (nm)	Primary decay time (μs)	Afterglow (% at 3 ms)
CsI(Tl)	4.52	6.1	54,000/0.02	550	1	0.5
$CdWO_4$ (CWO)	7.9	2.6	28,000/−0.3	495	2, 15	0.05
$Gd_2O_3:Eu^{+3}$	7.55	2.6	–/–	610	–	–
$(Y,Gd)_2O_3$:Eu, Pr, Tb (YGO)	5.9	6.1	42,000/0.04	610	1000	5
Gd_2O_2S:Pr,Ce,F (GOS)	7.34	2.9	50,000/−0.6	520	2.4	<0.1
Gd_2O_2S:Tb(Ce) (GOS)	7.34	2.9	50,000/−0.6	550	600	0.6
La_2HfO_7:Ti	7.9	2.8	13,000/–	475	10	–
$Gd_3Ga_5O_{12}$:Cr,Ce	7.09	4.5	39,000/–	730	150	<0.1

20.2.2.2 Photodetectors

The visible or near infrared light produced by the X-ray absorption in the scintillating screen is converted into an electronic signal by a solid state photodetector usually in the form of an array of silicon p-i-n photodiodes. For CT applications the photodiode must satisfy the following requirements:

- High quantum and geometric efficiency to improve the signal statistics.
- High shunt resistance. This reduces the offset drift of the detector system due to the variations of the photodiode leakage current caused by temperature changes. Typically, these changes create image artifacts at high attenuation levels.
- Low capacitance to reduce the electronic noise.
- Ability to connect a large number of the photodiode pixels to the data acquisition system. This becomes increasingly difficult for 32 or 64 slice CT scanners.

The majority of the 16 slice scanners use conventional front-illuminated (FIP) silicon p-i-n photodiodes technology (Fig. 20.5a). However, it requires electrical strips on the front surface and electrical wirebonds from the edges of the silicon chip to the substrate. When the number of slices approaches 64, the increasing number of conductive strips the active area of the channels becomes unacceptably small, and the high density of the wirebonds cannot be handled by conventional

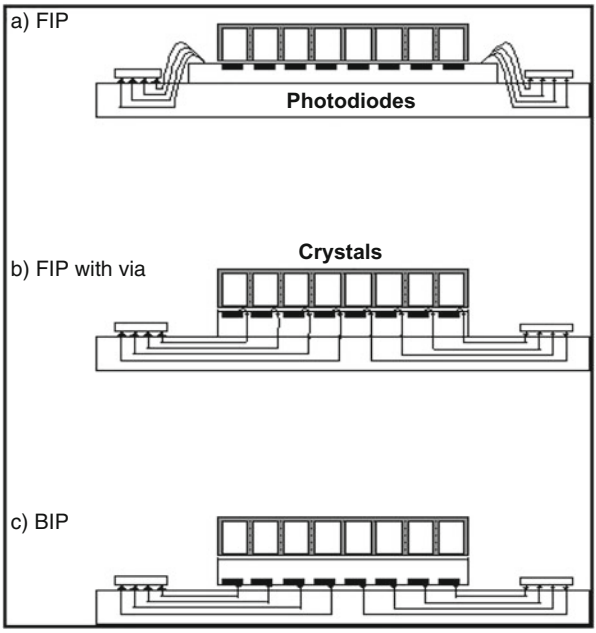


Fig. 20.5 Three types of photodiode arrays used in multislice CT. Front-illuminated FIP (**a**), Front-illuminated with via (**b**), and Back-illuminated BIP (**c**) (Courtesy R. Deych)

wirebond technology. In previous generation scanners, this limitation was addressed by combining the FIP technology with "vias" (electrically conducting feedthroughs) in the photodiode substrate [10]. The anodes of the photodiode elements were still wirebonded to via conductors, but the density of the wirebond was greatly reduced, because they were distributed over the area of the chip. The vias provided back-contacts for flip-chip connection to the detector board (Fig. 20.5b). Another approach is to use back-illuminated photodiodes (BIP) [11]. This solution solves the connectivity problem and the filling factor is almost 100% (Fig. 20.5c). It requires however very high resistivity silicon with large carrier lifetimes causing significant channel to channel cross-talk when standard silicon technology is used. In order to solve this problem BIPs are manufactured on 30 μm thick silicon wafer. Thinned BIP achieves almost 100% internal quantum efficiency in the spectral range 400–800 nm, and less than 1% cross-talk for 1×1 mm^2 pixels [12].

20.2.3 Scanner Geometry and Operating Conditions

20.2.3.1 Principle of Computed Tomography

The Computed Tomography principle introduced in the late 1960s by Allan MacLead Cormack and Sir Godfrey Newbold Hounsfield (Nobel Price laureates in Physiology and Medicine in 1979) marked a revolution in medical image reconstruction techniques. It is based on the relationship between the projections of a given parameter (X-ray attenuation for CT or radiotracer concentration for PET) integrated along line of responses (LOR's) at different angles through the patient and the Fourier transform of this parameter value distribution in the patient's body.

For the case of parallel beam illumination a projection at angle Φ is defined by the integration along all the parallel LORs of the parameter of interest as shown in Fig. 20.6 for a two-dimensional object f(x,y). The profile of all the LOR integrated values as a function of s, the radial distance from the centre of the projection, defines the projection at this angle Φ. The collection of all projections for $0 \leq \Phi < 2\pi$ forms a two-dimensional function of s and Φ. This function is called a sinogram because a fixed point in the f(x.y) object will trace a sinusoidal path in the projection space as shown in Fig. 20.6. A complex object will be represented in the projection space by the superposition of all the sinusoids associated to each individual point of the object. The line-integral transform of

$$f(x, y) \rightarrow p(s, \Phi)$$

is called the X-ray transform, also called the Radon transform for the two-dimensional case [13]. In this case, the projections are formed through a single transverse slice of the patient. By repeating this procedure through multiple axial slices, each displaced by a small increment in z, one can form a three-

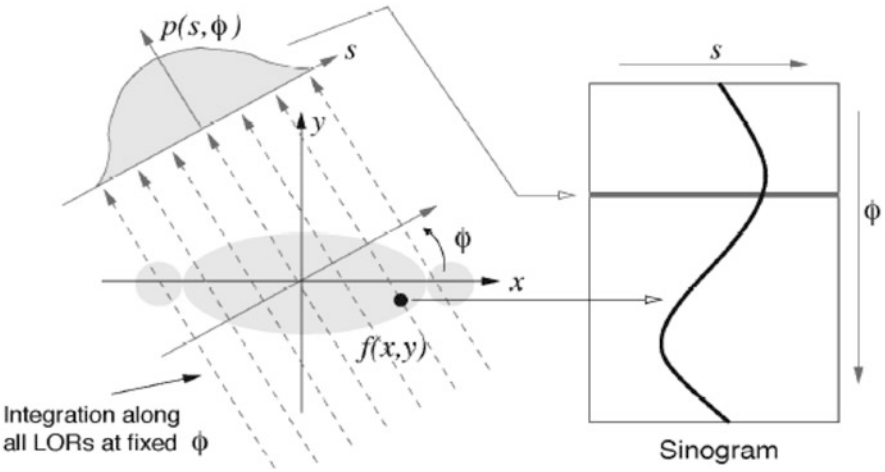


Fig. 20.6 Definition of the X-ray transform and the sinogram

dimensional image of a volumetric object f(x,y,z). It must be noticed that direct three-dimensional acquisition can be made by integrating LOR's not only in the transverse but also in oblique planes. Although more demanding in terms of computing power this fully three-dimensional approach is increasingly used in nuclear imaging (PET and SPECT), because it allows a significant gain in sensitivity.

The image reconstruction is based on the central-section theorem. This fundamental relationship in analytical image reconstruction states that the Fourier transform of a one-dimensional projection at angle Φ is equivalent to a section at the same angle through the centre of the two-dimensional Fourier transform of the object [14]. This is depicted in Fig. 20.7.

The image reconstruction process consists then in back-projecting and superposing all the data taken at all projection angles. However, to avoid oversampling in the centre of the Fourier transform (each projection will contribute to the central point, but increasingly less with increasing radial distance in the Fourier plane), the data are weighted, or filtered to correct for this oversampling. Basically, the Fourier transform of the back-projected image must be weighted by a cone filter

$$v = \sqrt{v_x^2 + v_y^2}$$

to decrease the values in the centre and increase them at the edges of the Fourier space.

However, in the specific case of X-ray CT the X-ray source is quasi-pointlike and the LOR's are not parallel. There is an important difference in the way parallel beam and divergent beam projections are back-projected. In a single view of divergent

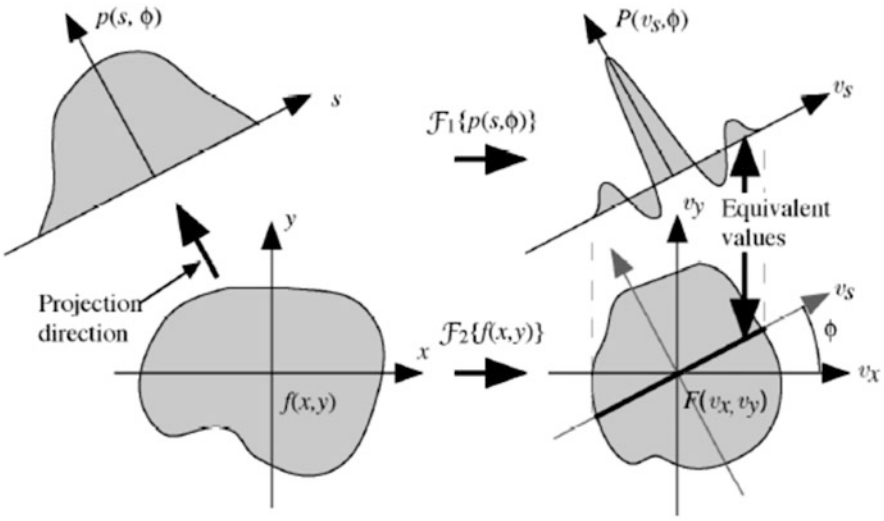


Fig. 20.7 Central-section theorem. $\mathcal{F}\{\ f\ (x, y)\ \}$ is the two-dimensional Fourier transform of the image and v_x is the Fourier conjugate of x

beam projections, the shift invariance of the image object is lost. As a consequence, equal weighting is not appropriate for back-projecting the measured divergent beam projections as it is in the parallel-beam cases. However, one can exploit the feature that in each single view, all the back-projections converge at the same X-ray focal spot. Therefore, the back-projection operation has a physical meaning only in a semi-infinite line: from the X-ray source position to infinity. Intuitively, an appropriate weight for the divergent beam back-projection operation should be a function of the distance from the X-ray source position to the back-projected point. If the distance from an X-ray source position $\overrightarrow{y}(t)$ to a back-projected point $\overrightarrow{x}$ is denoted as r, i.e. $r = \left|\overrightarrow{x} - \overrightarrow{y}(t)\right|$, then a weighting function $w(r)$ can be assigned for back-projecting the divergent beam projections. Using this general form of weighting function, a weighted back-projection can be defined as

$$G_d\left[\overrightarrow{x}, \overrightarrow{y}(t)\right] = w\left(r = \left|\overrightarrow{x} - \overrightarrow{y}(t)\right|\right) g_d\left[\hat{r} = \frac{\overrightarrow{x} - \overrightarrow{y}(t)}{\left|\overrightarrow{x} - \overrightarrow{y}(t)\right|}, \overrightarrow{y}(t)\right]$$

A measured projection value $g_d\left(\hat{r}, \overrightarrow{y}\right)$ is multiplied by a weight $w\left(r = \left|\overrightarrow{x} - \overrightarrow{y}(t)\right|\right)$ and then back-projected along the direction $\hat{r}$ to a point $\overrightarrow{x}$ with distance $r = \left|\overrightarrow{x} - \overrightarrow{y}(t)\right|$ as shown on Fig. 20.8.

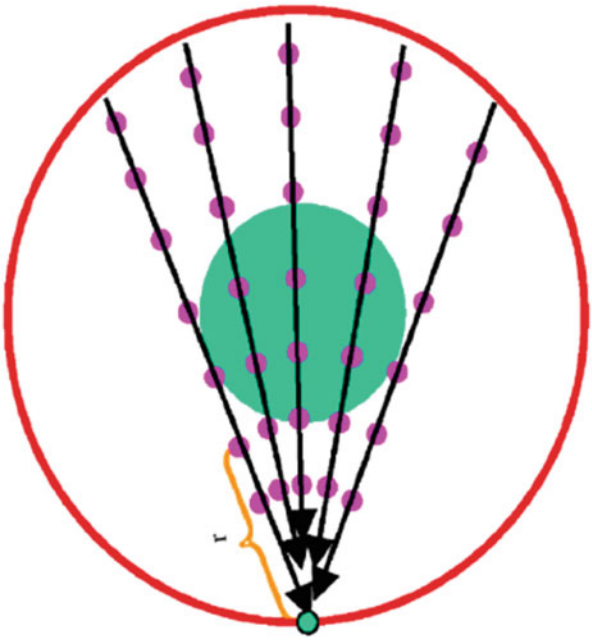


Fig. 20.8 Schematic representation of the divergent-beam weighted back-projection: The magenta points at seven different positions on each of the represented LOR's are visual guides to indicate how the sampling fraction varies as a function of the distance to the X-ray source, justifying the need for a distance-dependent weighting function in the backprojection algorithm (from [15])

20.2.3.2 Design of Modern CT Scanners

The most recent advance in CT scanning is the introduction of multi-slice helical scanning, sometimes known as spiral CT. A volume of tissue, e.g. the thorax or abdomen, is scanned by moving the patient continuously through the gantry of the scanner while the X-ray tube and detectors rotate continuously. The multi-slice systems offer the advantage over single-slice systems of being able to acquire information about the same volume in a shorter time, or alternatively to scan larger volumes in the same time or scan the same volume but obtaining thinner slices for better z-axis resolution. Helical CT has improved over the past years with faster gantry rotation, more powerful X-ray tubes, and improved interpolation algorithms [16].

The introduction of multi ring detectors and cone beam reconstruction algorithms have enabled the simultaneous acquisition of multiple slices: 4 slices in 1998, 16 slices 3 years later, 64 slices at the end of 2004, and up to 128 slices for the last generation scanners. Coupled with continuous increase of the gantry rotational speed (1.5 rotations per second in 1998, about 3 rotations per second in 2004) multislice acquisition is allowing shorter scan times (important for trauma patients, patients with limited ability to cooperate, pediatric cases and CT angiography),

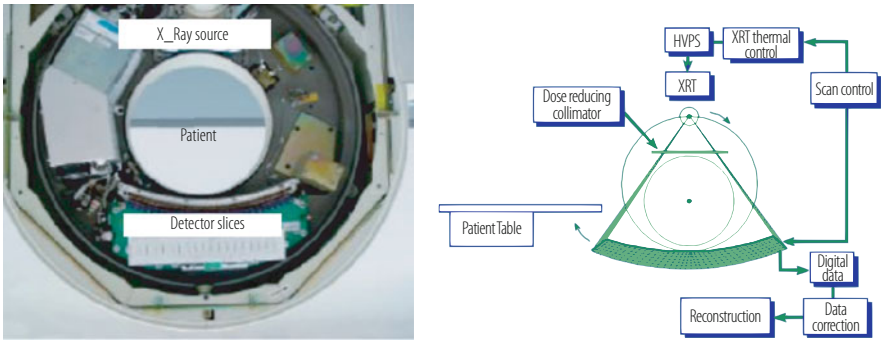


Fig. 20.9 Siemens Somaton 64 without cover and multislice CT block-diagram

extended longitudinal scan range (for combined chest abdomen scans, such as in oncological staging) and/or improved longitudinal resolution (typically 0.5 mm per slice). It has further improved the performance of the existing applications such as angiography and detection of lung and liver lesions as well as paved the way to the introduction of new ones, most notably in cardiology, where high quality images can be obtained in 10–20 heartbeats or in a single breath-hold only.

A third-generation multislice CT scanner and a block-diagram is shown in Fig. 20.9. The slip-ring technology was introduced at the end of 1980s and allowed the spiral scanning mode, when the X-ray tube and 2D arc of multislice detector system are rotated continuously around the patient while the scanning table is translated through the rotating gantry. The scan parameters, selected by the radiologist, define the X-ray tube protocol, X-ray collimation, patient table motion, data acquisition, and reconstruction parameters.

The scanner can be designed for a fixed slice collimation. It is however desirable, although more challenging, to design the detector in such a way as to meet the clinical requirement of different slice collimations adjustable to the diagnostic needs. There are basically two different approaches, the matrix detector with elements of a fixed size or the adaptive array principle (Fig. 20.10). As shown in the figure for a 16 slice array in an expanded way the cone beam geometry introduces a smearing over the field of view, which Increases the slice thickness on the edges of the cone compared to the centre.

There is therefore no need to have the same number of detector elements in the centre and in the periphery of the detector.

Most of the modern CT scanners have multiple, up to 128, detector rings, or slices. Typical CT scanners have a field of view FOV = 50 cm, and a spatial resolution of 0.5 mm in the middle of the FOV. Therefore, each detector ring houses more than 1000 detector elements per slice. The electronic channels amplify and filter the detector current and measure the filtered current at small time interval, called "view time" T_w, which is the time in which the disc rotates approximately one 1/4 to 1/3 of a degree. Since the rotational speed of modern CT scanners

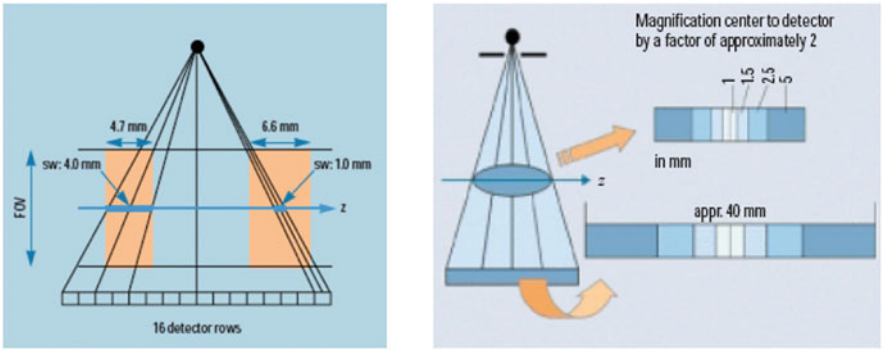


Fig. 20.10 Principle of the matrix detector (left) and of the Adaptive Array Detector (right)

can be as high as 3 or even 4 rotations per seconds, T_w can be as short as about 250 μs. At a typical focal spot-detector distance of 1000 mm, the exposure rate at the detector can reach ~0.1 Gy/s. Single detector channel detects up to 300,000 X-ray photons per sample at 3000 Hz acquisition rate. At such high photon fluxes, the detector cannot be operated in counting mode, and the majority of medical X-ray CT scanners operates in current measurement or integration mode.

One of the most widely used CT scanner, the Siemens SOMATOM Sensation 64 [17], is using a scintillating ceramics detector head. The X-ray focal spot is switched in two different positions during a view time, to reduce the aliasing of sampled data in the translational direction, i.e. along the z-axis. Consequently, the readout electronics must sample and measure the input signal two times in each view time. This machine uses an adaptive array detector with 40 detector rows in the longitudinal direction. The 32 central rows have a slice width of 0.6 mm in the centre of the field of view, whereas four detector rows on each side (in the penumbra of the collimated X-ray source) have a slice width of 1.2 mm. The slice widths being determined at the isocentre the actual detector size is about twice as large, due to the geometrical magnification (Fig. 20.11). Acquisition of 64 slices per rotation is possible through the use of a special X-ray tube with a flying spot capability (Fig. 20.12). The electron focal spot is wobbled between two different positions of a tilted anode plate by an variable electromagnetic field, resulting in a motion of the X-ray beam in the longitudinal direction. The amplitude of this periodic motion is adjusted in such a way that two subsequent readings are shifted by half a slice width in the longitudinal direction.

In general, it should be remembered that the performance factors of image quality, dose and speed can each only be improved at the expense of the other parameters. High contrast resolution in the final image is affected by noise, matrix size and contrast. Low contrast detection is affected by the size of the object, windowing and image noise. Image noise itself is affected by exposure factors, detection efficiency, slice width and, most critically, by the algorithms used in the reconstructions.

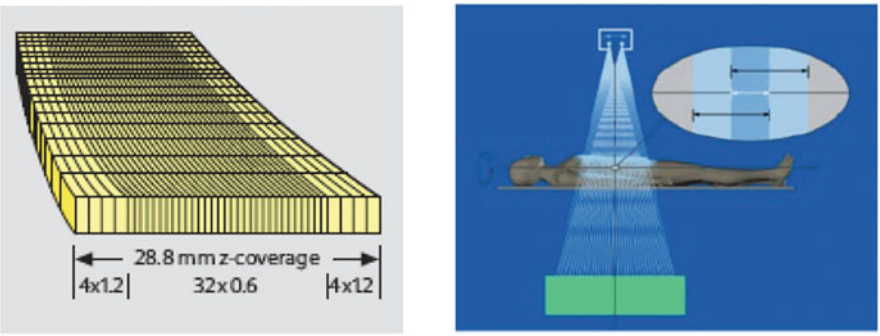


Fig. 20.11 Adaptive array detector with 32 slices of 0.6 mm in the central part, resulting in 64 slices with 0.3 mm sampling at the isocentre (Courtesy Siemens Medical Solutions)

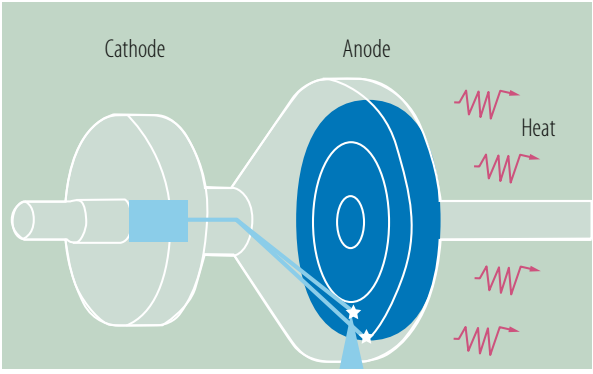


Fig. 20.12 Schematic drawing of a rotating envelope X-ray tube (Siemens STRATON, Forchheim, Germany) with z-flying spot technique (Courtesy Siemens Medical Solutions)

It must be noted that for modern scanners the X-ray tube operates with a higher duty cycle: heat output and heat dissipation are therefore a concern in the design of such multislice CT-scanners.

Another important trade-off is related to the radiation exposure of the patient. The continuing quest for better spatial resolution imposes ever smaller detector sizes. As the area of each detector cell decreases, the amount of X-rays incident on the detector decreases, leading to an increase in statistical noise. Retaining the original signal/noise ratio (and therefore the same level of contrast detection power) requires an increase in the number of X-rays and hence patient radiation dose. There is, therefore, a balance to be struck between radiation dose and resolution.

Recent developments are aiming at further decreasing the scan time and, most importantly, reducing the dose exposure to the patient. This is achieved by the introduction of the dual X-ray technology, with two X-ray sources of different energy (typically 80 kV and 140 kV) and selective photon shield for better spectral separation. Combined with two 128 slices detector panels and a

rotation speed of 0.28 s, the SOMATOM Definition Flash Spiral from Siemens achieves ultrafast image acquisition (not necessitating a breath hold) with an dose exposure approaching 1 mSv for a number of protocols, to be compared to 20 mSv 10 years ago (https://www.healthcare.siemens.com/computed-tomography/dual-source-ct/somatom-definition-flash/technical-specifications). A summary of the recent progress and new trends in CT imaging can be found in [18, 19].

20.2.4 Future of X-Ray Imaging

X-ray imaging is the historical imaging modality since the discovery of X-rays and the pioneering work of W. Roentgen in 1895. It is still the most widely used imaging diagnostic tool for physicians with nearly half a billion X-ray exams performed every year worldwide.

One major thrust for the future X-ray imaging devices is to obtain higher resolution data at a faster rate. For instance, cardiac applications would substantially benefit from CT scanners able to acquire heart images in one heartbeat or less so that motion artifacts can be minimized. One direction being pursued are scanners featuring multiple X-ray tube/data acquisition combinations operating simultaneously. Moreover, tubes incorporating a smaller focal spot are being introduced, enabling higher spatial resolution (up to around 25 line pairs per cm). Various other medical applications, such as surgical ones or the rapidly developing interventional radiography, would substantially benefit from CT scanners able to perform acquisitions at both normal and ultra-high spatial resolutions, as those required in fluoroscopic procedures. Ultra-high resolution can be achieved by substantially shrinking the physical size of both the focal spot and detector elements. Similar trends to reduce the pixel dimensions are observed for two-dimensional detector arrays used in digital radiography. This results in a considerable increase in the number of pixels and increases the complexity of the acquisition system in CT and planar digital radiography (see Sects. 20.2.4.1 and 20.2.4.2).

Present X-ray imaging only provides morphologic information but no information about the physiology of the organs under examination. However ongoing research suggests that information about the pathology of a tissue is conveyed not only by its overall X-ray attenuation, but also by its selective absorption at different X-ray beam wavelengths. This opens a new and exciting field: exploiting the new singly photon counting techniques for studying tissue pathologies with X-ray spectral images (see Sects. 20.2.4.3 and 20.2.4.4).

20.2.4.1 Indirect Detection with Phosphor Screen

The choice of the scintillating material is of course the key for a higher segmentation of a new generation of X-ray devices, as the pixel size is determined by mechanical properties of the crystal like hardness, cleavage, mechanical processing yield

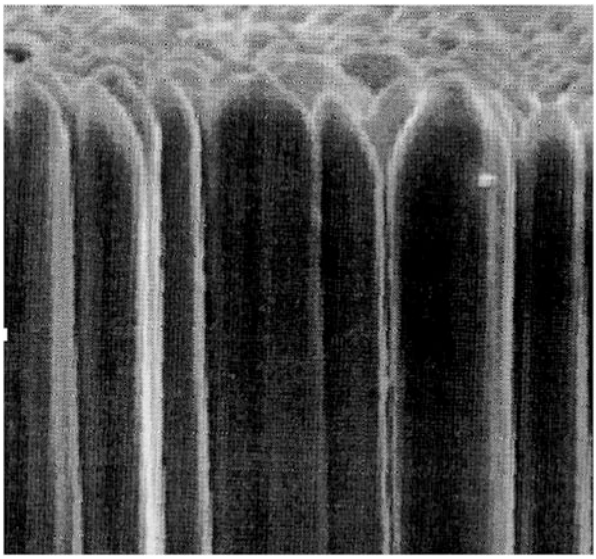


Fig. 20.13 Column structure of vapor deposited CsI(Tl). Columns have a typical diameter of 10 μm and a length of 500 μm

and cost. Large efforts have been devoted recently on specific technologies to develop a solid-state dynamic X-ray sensor with digital readout for matrixes manufacturing with sub-millimeter resolution. So called columnar structure screens were developed [20]. The rapid progress on position sensitive photomultipliers (PSPMT), Silicon photodiodes with different designs and Geiger mode Silicon photomultipliers (SiPMT) open attractive possibilities for pixel based arrays. The current design is based on large a-Si photodiodes (substrate) coupled to a CsI(Tl) layer. The scintillator layer growth is nucleated on the pattern substrate and transferred to a columnar system separated with grain boundaries as seen in Fig. 20.13. Each CsI(Tl) column is not only a scintillation pixel but also a light guide. This guide prevents or at least strongly suppresses the radial light spread and might be the way to obtain very high spatial resolution. Columnar structure growth technique allows to get 3–5 μm diameter columns and the pixel size is defined by the Si pad size as seen from Fig. 20.14. Currently, flat panels with dimensions up to 40×40 cm^2 are developed to image the human chest.

It should be noted that it is possible to use non-pixilated screens for low energy X-rays. If X-rays are absorbed in a very thin crystal layer, the angle of the emitted light is small (for the thin film) and the crosstalk to the neighbor photo-receiver is negligible maintaining therefore a good spatial resolution. The search for materials for such applications is now of very high importance.

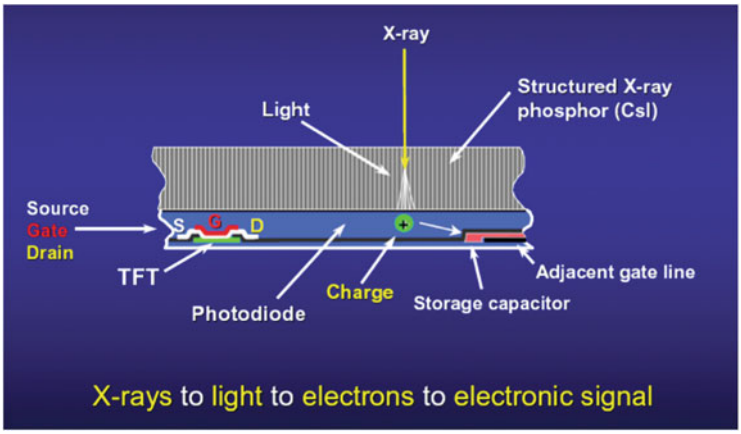


Fig. 20.14 Integrated columnar CsI(Tl) and a-Si photodiode readout for new generation X-ray flat panel (Courtesy J.A. Seibert, UC Davis Medical Centre, CA, USA)

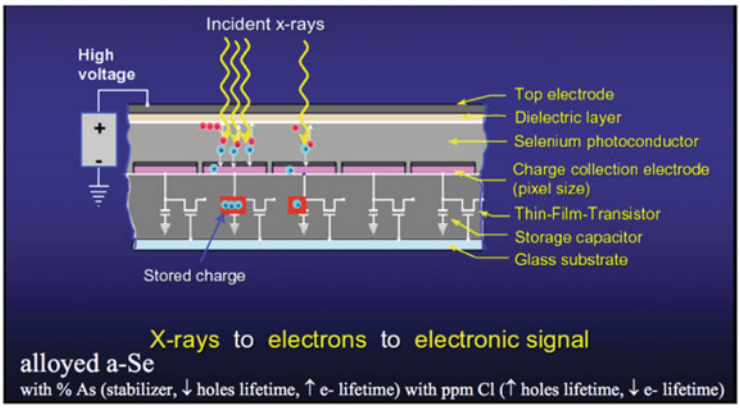


Fig. 20.15 Direct conversion detector for new generation X-ray flat panel (Courtesy J.A. Seibert, UC Davis Medical Centre, CA, USA)

20.2.4.2 Direct Conversion Screen

Another even more radical departure from the present X-ray detector technology may be the use of high-density room temperature semiconductors.

As shown in Fig. 20.15, direct detection flat panel technology is based on a uniform layer of X-ray sensitive photoconductor, e.g., amorphous selenium (a-Se) to directly convert incident X-rays to charge, which is subsequently electronically read out by a two-dimensional array of Thin-Film-Transistors (TFT). During readout, the scanning control circuit generates pulses to turn on the TFTs one row at a time, and transfers the image charge from the pixel to external charge sensitive amplifiers. They are shared by all the pixels in the same column. Each row of the detector

typically takes about 20 μs to read out. Hence a detector with 1000 × 1000 pixels can be read out in real-time (i.e., 30 frames/s).

A challenge for this approach is the practical implementation of the complex pixel design over a large area with consistent and uniform imaging performance. The problem of charge collection efficiency and speed for materials with high Z and sufficient thickness remains a major concern. Substantial technical problems must be resolved before these technologies will be implemented in commercial X-ray devices.

20.2.4.3 Single Photon Counting

The impressive CT images shown in the literature (Fig. 20.4) require several tens to 100 times higher X-ray exposures compared to standard radiography (typically 20–50 mSv as compared to 0.1 mSv). On the other hand, the enormous research effort on particle detectors has led to the development of digital X-ray detectors with very small pixels, based on silicon (Medipix) [21] and on gaseous detectors (GEM, Micromegas) [22]. The major and unique feature of these devices is their capability to work in single photon counting mode up to very high rates. Excellent high contrast images can therefore be obtained with X-ray doses up to 10 times smaller than for standard X-ray systems working in current mode.

There are a number of advantages of counting systems over current mode systems, such as:

- maximization of the contrast resolution, limited by the intrinsic Poisson statistics of the number of detected photons
- elimination of the excess noise resulting from the variance in the number of visible photons produced by the X-ray conversion in the phosphor screen, also called Swank factor [23]
- linear behavior over the whole dynamic range, which can be adapted to the specific application requirements
- possibility of implementing multiple thresholds for energy discriminating techniques, which can be used for instance for dual energy radiography, K-edge subtraction or Compton scattering discrimination
- no need for an energy dependent weighting factor as each event has equal weight whatever its energy.

This results in much better image contrast performance and significantly lower doses as shown in Fig. 20.16 in a comparative study of the signal-to-noise ratio for a 2 mm thick tumour as a function of the X-ray tube current for different mammography systems. The Medipix single photon counting device achieves the same image quality as the best commercial mammograph working in current mode for typically half the dose.

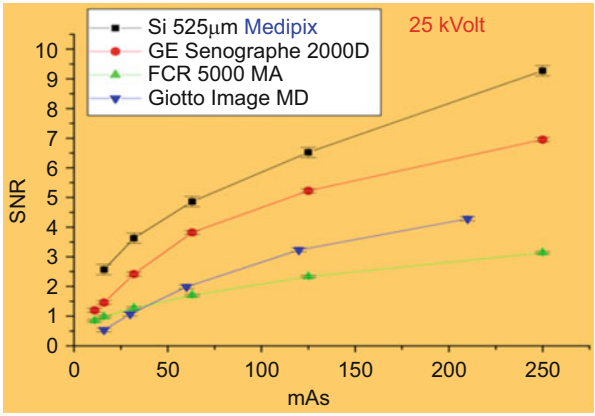


Fig. 20.16 Comparative study of the signal to noise ratio of different mammography systems as a function of the anodic current times exposure time, which is proportional to the X-ray fluence (from [24])

20.2.4.4 Spectral X-Ray Imaging

The introduction of hybrid pixel detectors in X-ray imaging**,** where the sensor array and the matching read-out chip are processed independently and are connected together only in the final step, has allowed high dynamic, noise free images to be recorded on the basis of single photon counting techniques [25]. Moreover, among the most promising recent developments in CT is the use of spectral information to improve contrast discrimination, by acquiring data with different energy thresholds. In traditional CT imaging, the overall attenuation of X-ray intensity is measured by the detector, but the detected X-rays are not spectrally resolved. This introduces a bias in the images because the absorption of X-rays by different materials depends on the X-ray energy. A significant amount of information can therefore be gained by including spectral data in the CT reconstruction process. Based on differences in X-ray absorption, different materials can be distinguished and quantified with a single spectral CT scan.

Two principal methods provide spectral CT data. The first method, dual-energy (DE) CT, uses X-ray sources with two different energy spectra and energy integrating X-ray detectors. The second method uses a single X-ray source but has energy-resolving detectors (photon counting detectors) that measure the energy of each detected photons. DE CT is currently used clinically and has been successful in improving imaging for a variety of applications.

While the soft tissue attenuation coefficient is rather wavelength independent, the photoelectric effect in high atomic Z materials strongly depends on X-ray energy. This feature can be exploited by using contrast agents containing such high Z elements. The attenuation coefficient of such substances (calcium in bone, iodine, gold) will show significant differences if the two energy spectra are recorded on either side of the K-edge for these heavy elements. The large increase in attenuation at energies above the K-edge leads to large signal differences between the two scans. By combining data from the two energy sets, these high Z materials can be distinguished and quantified. Figure 20.17 shows scans of a mice after injection

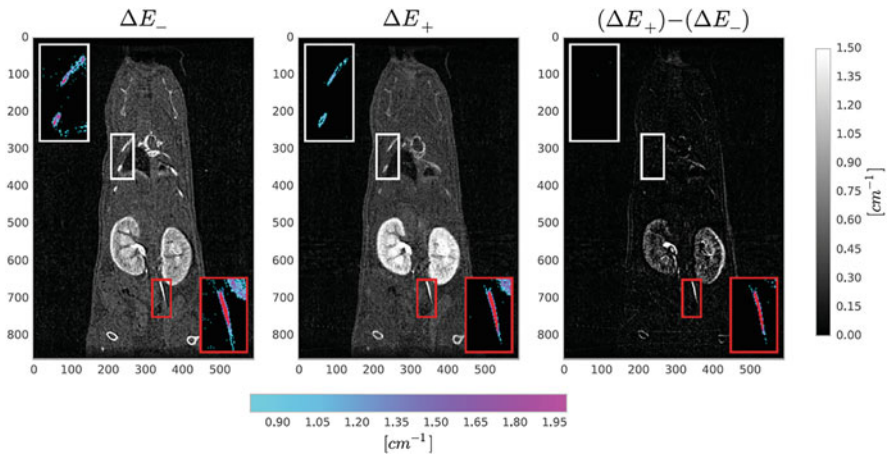


Fig. 20.17 Coronal slices of a sacrificed mouse after injection of an iodinated contrast agent. From left to right: slice reconstructed in the energy window just below the iodine K-edge, in the energy window just above the iodine K-edge and K-edge image (subtraction of the first slice from the second one). Zooms of two ribs and the ureter are also shown (from [26])

of a iodinated contrast agent. The difference between the scan taken in an energy window just below the iodine K-edge (left) and the scan taken in an energy window just above the K-edge (centre) shows the iodine concentration in some parts of the kidneys. More information can be found in [26].

20.3 Single Photon (SPECT) and Positron (PET) Emission Tomography

20.3.1 SPECT and PET Working Principle

Nuclear medicine relies on using radioactive molecules administered to a patient for diagnostic or therapeutic purposes. Radioactive molecules behave in vivo the same way as their non-radioactive "natural" equivalent involved in the metabolic or molecular processes under study. Nuclear medicine is used daily in oncology, cardiology, neurology, paediatrics, rheumatology or orthopaedics for diagnosis and therapy.

A new and recent concept is molecular imaging. It provides the ability to visualize and quantitatively measure in-vivo the activity of different biological and cellular processes activated or depressed in some pathologies.

The working principle of emission tomography is to image γ rays emitted by the radiotracers injected into the patient. Contrary to X-ray CT and standard nuclear magnetic resonance, which provide very precise images of the anatomy of

organs, nuclear molecular imaging modalities give in vivo access to the quantitative functioning of these organs.

20.3.1.1 SPECT

In Single Photon Emission Computed Tomography (SPECT) a molecule involved in the metabolism of the patient is labeled by a single photon emitter (usually ^{99}Tc emitting one 140 keV γ ray). After injection, this molecule concentrates preferentially in the organs or tumours where this metabolic function is active and allows their imaging through the reconstruction of the γ ray emitting points. The most popular technique is based on the "Anger logic", where γ rays are directed through a multi-hole collimator to a large slab of Sodium Iodide (NaI) or Cesium Iodide (CsI) scintillator. The coordinates of the interaction point are then determined by comparing the signals from a set of photomultipliers (PMT) coupled to the crystal, by the centre of gravity method (Fig. 20.18). This technique, called scintigraphy, is still largely used in many hospitals and medical imaging labs, but suffers from a relatively poor space resolution, of the order of a few centimeters.

More recent detector designs are based on discrete scintillating pixels coupled to multichannel photodetectors, such as multianode photomultipliers or avalanche photodiode matrices. But the most impressive progress has been made on the collimator, which is the main limiting factor for the spatial resolution and the sensitivity of SPECT devices. Several configurations have been studied:

- The parallel collimator, in which all the septa are perpendicular to the crystal surface
- The slanthole collimator, where the holes are parallel to each other but slanted, all in the same direction
- The fan beam collimator, where the holes are focused to a line
- The cone beam collimator, where the holes are focused to a point
- The pinhole collimator, at some distance from the crystal, where the field of view increases with the distance from the object

The best results so far are achieved with multi-pinhole configurations, for which sub-millimeter spatial resolution and sensitivities at the level of 1 cps/KBq have been obtained on small animal imaging SPECT cameras. The counterpart of using collimators, and particularly pinholes, is a reduction of the overall sensitivity of the SPECT camera. For clinical applications a compromise needs to be found between sensitivity and spatial resolution.

The spatial resolution is usually given by the Full Width at Half Maximum (FWHM) of the so-called point spread function (PSF):

$$FWHM = \frac{D}{L}(z_0 + L + B)$$

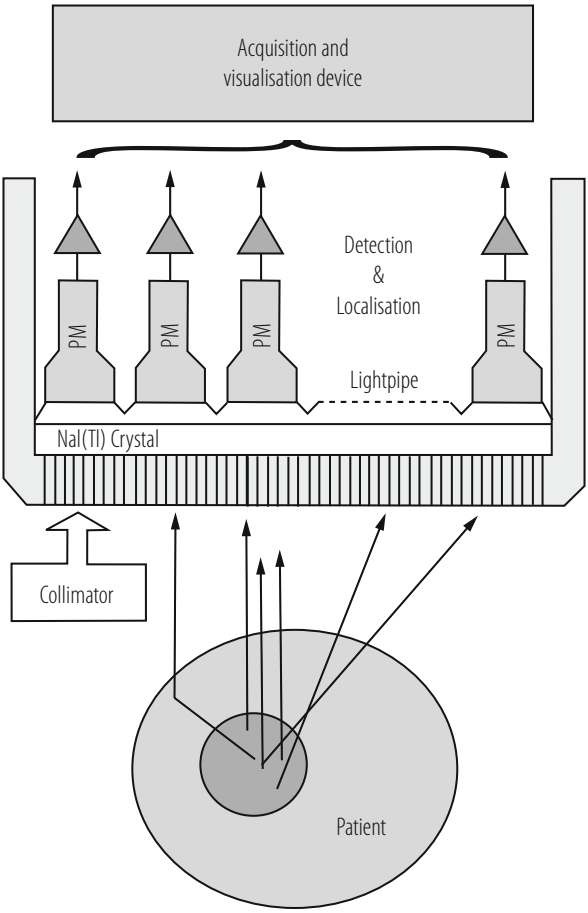


Fig. 20.18 Principle of an Anger camera

where D and L are the collimator hole diameter and length, z_0 is the distance from the γ-ray source to the collimator entrance and B is the distance between the collimator back face and the image plane in the crystal.

The collimator introduces an important loss of efficiency by a factor of 10^3 to 10^4. The resulting efficiency is given by the following formula:

$$\eta = \varepsilon \left(\frac{a_{hole}}{a_{cell}} \right) \left(\frac{a_{hole}}{4\pi L^2} \right)$$

where a_{hole} and a_{cell} are the collimator hole and cell area respectively and ε is the γ detection efficiency in the scintillator.

20.3.1.2 PET

In the case of Positron Emission Tomography (PET) the functional molecules are labeled with a β^+ emitter generally produced in a cyclotron. These PET tracers, injected into the patient, simulate natural sugars, proteins, water and oxygen presence, circulation and accumulation in the human body. Once fixed in the organ or the tumour, the molecule emits positrons, which annihilate very quickly on contact with the tissue, emitting two gamma photons located on the same axis—called the line of response (LOR)—but in opposite directions, with a precise energy of 511 keV each (Fig. 20.19).

The coincidence detection scheme introduces therefore an electronic collimation, which greatly enhances the background rejection as compared to SPECT. Moreover, the line of interaction being precisely determined by the two detectors hit in coincidence, there is no need for a collimator system, which severely reduces the sensitivity of SPECT cameras. In order to simplify and reduce the image reconstruction time the first generation PET scanners used septa to restrict the acquisition to transversal slices through the patient in a so called 2D acquisition mode. The slices were then combined off line for a 3D image reconstruction. Modern PET scanners benefit from the considerable progress in computer power and directly acquire data in 3D mode without septa (i.e. recording all the LORs independent of their direction relative to the scanner axis), which results in a significant gain in sensitivity.

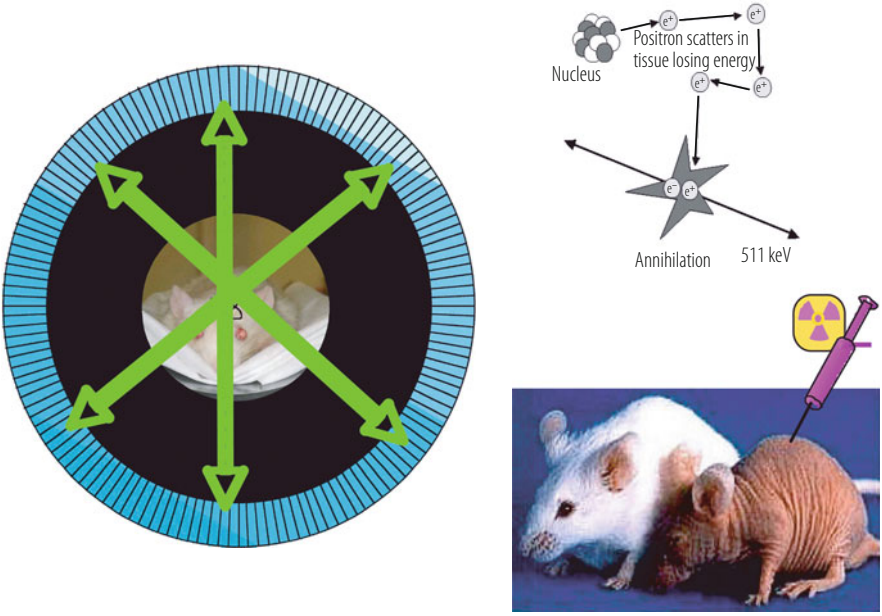


Fig. 20.19 Principle of a PET scanner

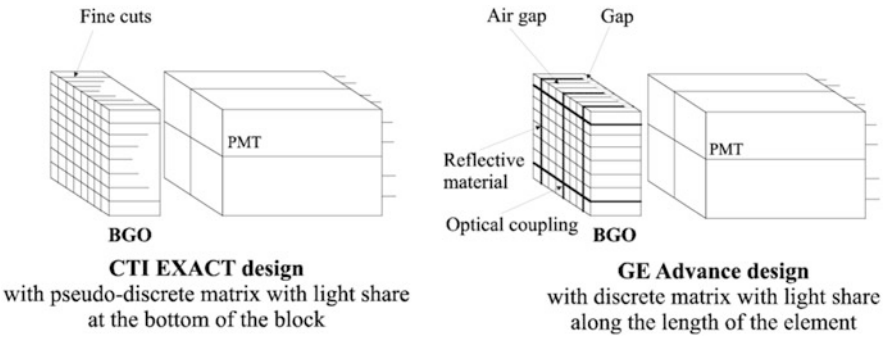


Fig. 20.20 The readout quadrant-sharing scheme for the CTI (now Siemens) and for the General Electrics PET scanners

Until recently, as a result of a compromise between performance and cost, PET scanners were using partially segmented BGO crystals readout by groups of four PMT's (quadrant-sharing scheme), allowing a reconstruction precision of the order of 4–5 mm (Fig. 20.20). Modern machines are going progressively to higher segmentation of the crystals and of the readout to achieve higher spatial resolution. Resolutions of the order of 1 mm have been reached at least for small dedicated machines, such as breast imaging devices or for small animals.

Positron Emission Tomography measures the uptake of the tracer in different organs or tumours and generates an image of cellular biological activities with a much higher sensitivity than any other functional imaging modality. The PET images can be used to quantitatively measure many processes, including sugar metabolism, blood flow and perfusion, oxygen consumption etc. Moreover, specialized PET scanners designed for experimental small animal studies (mouse, rat, rabbit) are powerful tools for fundamental research in disease models, new therapeutic approaches and pharmacological developments. The most commonly used radio-isotopes are ^{18}F with a lifetime of 109.8 min, ^{11}C (20.4 min), ^{13}N (10 min) and ^{15}O (2.1 min), the last three ones being among the basic building blocks of organic systems and therefore being easily introduced chemically in molecules involved in metabolic or pharmaceutical reactions. A typical example is FDG (^{18}F labeled fluorodeoxyglucose), which allows monitoring the energetic consumption of the cells in different parts of the body. FDG is a glucose analog, where a hydroxyl group has been substituted by a ^{18}F atom. Once phosphorylated by the hexokinase enzyme into FDG-6-phosphate, it remains trapped in the cell, where it accumulates. The interest in labeled glucose lies in the fact that tumour cells are characterized by an increase of glycolisis and expression of glucose transporters, such as GLUT-1, as compared to healthy tissues. This increase of FDG metabolism allows detecting tumours and related metastases through their abnormally high glucose concentration and therefore increased γ-ray activity.

PET has a very high sensitivity, at the picomolar level, which makes it one of the essential tools of molecular imaging with applications in many areas such as

expression and occupancy of therapeutic molecular targets, pharmacokinetics and pharmacodynamics, mechanisms of therapeutic action and functional response to therapy.

20.3.2 Detector Challenges for Modern Nuclear Medicine

The spectacular development of in-vivo molecular imaging will allow in the near future bridging the gap between post-genomics research and physiology and opening interesting perspectives for new diagnostic and therapeutic strategies for many diseases. Nuclear medicine and particularly PET imaging are already playing and will play an increasing role in molecular imaging [27, 28]. Constant progress in the medical and biological fields implies that imaging performances have to be continuously improved. In order to fulfill the needs of quantitative cell and molecular imaging, of dynamic studies over a certain time and of individualized therapy focusing on the patient's genotype, major technical improvements [29] will be necessary, comparable to those in large particle detectors, in order to deal with:

- integration of a very large number of increasingly compact measuring channels (several hundred thousands)
- data acquisition rates at the level of tens gigabytes/second
- several billions of events to reconstruct an image
- about 1000 gigabytes of data per image and commensurate computational power for the reconstruction
- integration of multiple technologies requiring pluridisciplinary competences for complex, compact and reliable systems.

The challenge for functional isotopic imaging lays in its capacity to identify the specific molecular pathways in action in a metabolic process and to quantitatively measure their relative metabolic activity. To achieve this, it is necessary to improve both the imaging system's spatial resolution, that is, its capacity to discriminate two separate objects, and the measurement's signal to noise ratio, that is, how precisely a metabolic agent's concentration in a body area can be determined. The precision of the concentration measurement depends mainly, but not only, on the imaging system's sensitivity, and therefore its capacity to accumulate the statistics needed to tomographically reconstruct the radiopharmaceutical tracer distribution. Moreover, the location of this metabolic activity must be precisely associated to the organs or parts of the organs under examination. This explains the increasing demand for combining functional and anatomical imaging devices.

The perspectives to develop isotopic imaging with multimodality and multifunctional capability revolve around three goals:

- improving sensitivity
- improving spatial resolution
- improving temporal resolution

20.3.2.1 Improving Sensitivity and Specificity

Sensitivity is defined as the ratio of the detected number of radioactive decays and the radioactivity injected into the patient and fixed on the organ under study. It reaches at best 10% in the case of PET scans on small animals, and a few percent only in the case of whole body PET. The main losses arise from poor geometrical acceptance, gaps between crystals, rejection of Compton events due to partial conversion of the γ-ray in the crystal and electronics dead time. Moreover, whole body PET scanners visualize only the patient's thorax in one acquisition run, which is sometimes a limiting factor, for instance in oncological studies of bone metastases in limbs. In the last few years, the use of faster scintillating crystals and electronics and improved geometrical acceptance has allowed to reach the above-mentioned sensitivity levels.

However, sensitivity has to be further improved for several reasons. First, examination durations have to be shortened. Today, a whole body scan lasts between 10 and 20 min. It would be desirable to reduce this time to a few minutes to improve the patients' comfort. It would also increase the exploitation of costly equipment and infrastructures with a significant impact on the cost of examinations. Shorter acquisition times would also improve the image quality because the impact of the patient's natural movements—breathing and cardiac activity, digestive bolus, etc.—would be reduced. Quicker metabolic processes could be followed, which are crucial for pharmacokinetic studies.

There is a strong interplay between sensitivity and spatial resolution, since the signal-to-noise ratio per voxel is the relevant image quality factor. Doubling the linear spatial resolution (i.e. reducing the volume by a factor 8), requires a 16 times (because it requires two detector pixels to identify one voxel) higher noise equivalent rate if the statistical quality of the image is to be maintained. The acquisition time of an image depends on many factors, which all influence the noise-equivalent measurement of the imaging system and of the radioactivity administered to the patient. Some of the relevant parameters are: the imaging system's geometrical acceptance; the efficiency and energy resolution, improving the energy selection of events and hence discriminating in a coincidence system the diffused events; the time resolution allowing to reduce the width of the coincidence window and rejecting random coincidences more efficiently; dead time of the detectors.

Increasingly, specific molecular signatures for the major diseases are being evaluated to devise individually targeted therapies adapted to the patient's genotype. This requires ever more performing equipment and more specific protocols. Indeed, it is highly desirable to study different molecular pathways simultaneously and to record the intensity, the range, the localization and the temporal development of various biochemical processes in their natural environment in the human body. In this way, the nature of the pathology can be established, at least partially, through molecular imaging, using an array of radiopharmaceuticals giving information on cell proliferation (FLT), on energetic metabolism (FDG) or on aminoacid synthesis (methionine) in the various tissues. For multitracer analysis of various biochemical or pathophysiological processes several radioactive tracers have to be administered.

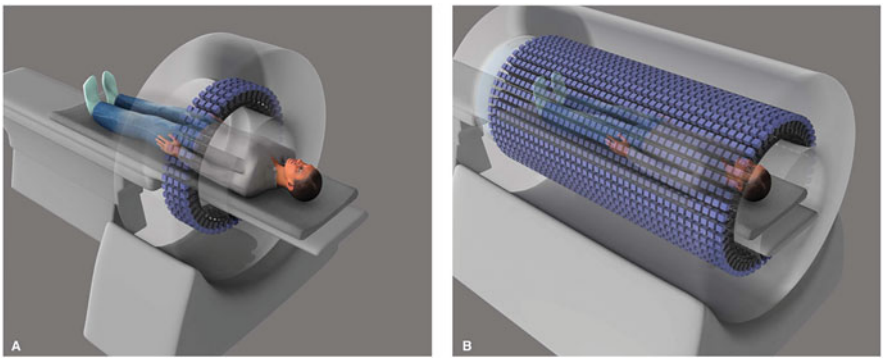


Fig. 20.21 Conventional PET scanner (**a**) and Total-Body PET (TB-PET) scanner (**b**). From [30]

In order to keep the doses tolerable for the patient, high-sensitivity PET scanners have to be developed, which would also open new prospects for young or pregnant women and for children.

The obvious approach for increasing the sensitivity is to increase the geometrical acceptance of the scanner. In some cases, developing dedicated equipment might be the right solution to study an organ (brain, breast, prostate) in a more efficient and optimized way. In this case the detectors can be placed closer to the organ under study, increasing therefore the geometrical acceptance of the events.

Until recently the length of the detector cylinder, or length of the system's sensitive volume ("field of view"), has remained essentially at the level of 15–18 cm (Fig. 20.21a), resulting in a very small geometric efficiency $\leq$0.2. Several static views need therefore to be acquired to perform a "whole body" scan, i.e. from the head to the pelvis. This procedure presents major limitations, as the acquisition time increases in proportion to the number of views. A "standard study" with the injection of 10 mCi of ^{18}F-FDG takes about 3 min per view and, therefore, approximately 20 min per entire study.

A few year ago, the EXPLORER consortium (http://explorer.ucdavis.edu) had launched the concept of a Total-Body PET scanner (TB-PET) to realize the full potential of PET—extending the FOV to cover the entire length of the body (Fig. 20.21b).

In TB-PET, the vast majority of the emitted photons could be captured. This step change in technological evolution would mean simultaneous coverage of all the tissues and organs in the body, with an overall >40-fold gain in effective sensitivity and a >6-fold increase in signal-to-noise ratio compared with whole-body imaging on current PET scanners. The challenge of funding the construction of the first prototype machine, an expensive novel device, was successfully overcome in September 2015 through funding from the NIH Transformative Research Award program, which recognizes high-risk, high-reward, paradigm-shifting innovative research [30].

To further improve sensitivity, we need denser and faster scintillating crystals or direct conversion materials, more compact and adaptable geometries, lower-noise and faster acquisition electronics, more parallelized acquisition architecture with integrated processing power, and at least partial use of the information included in the events diffused in the patient or in the detector. Potential progress in these fields are described in the next paragraphs.

Another promising way to significantly increase the sensitivity is to push the limits of time-of-flight PET scanners, as will be explained in Sect. 20.3.2.3.

20.3.2.2 Improving Spatial Resolution

Spatial resolution reaches 1.5–2 mm at the centre of the field of view of small animal PETs, but worsens off-axis. Modern technologies have allowed to reach a 1 mm resolution for small animals PETs or for scanners dedicated to specific organs, and a 4–5 mm resolution for whole body scanners.

Good spatial resolution is obviously of value for the study of small animals, but also for humans: increasingly smaller structures which are involved in specific metabolic processes can thus be visualized. Anatomical localization can also be more precise and combining CT or MRI information can be improved. But it is in the field of quantification that the improvement potential is likely to be the most significant. By reducing the blurring caused by insufficient spatial resolution (also called partial volume effect), the dynamic sensitivity of the radiotracer's concentration measurement, also called Standard Uptake Value (SUV), can be significantly improved.

There are four factors, which limit a PET camera's spatial resolution:

- the positron's mean free path: once the ligand has fixed itself on the organ or tumour being investigated, the radioisotope used to label it emits positrons with a kinetic energy depending on the isotope. As the annihilation probability of this positron is maximum when the positron has sufficiently slowed down, there is a difference between the positron emission point and its annihilation point. This difference is about 0.5 mm in the case of ^{18}F, but it can reach several mm for other isotopes—4.5 mm for ^{82}Rb, for instance. This blurring is often considered as an intrinsic limitation to the PET spatial resolution, but it can be significantly attenuated thanks to various electromagnetic artifices. For instance, the positron's trajectory usually revolves around the lines of a magnetic field—naturally present in the case of a combined PET-MRI camera—, which therefore reduces its conversion distance. This is however only effective in the plane perpendicular to the magnetic field and for new generation of MRI devices with high field (7T or more). It also has to be noted that positron annihilation probability as a function of its speed is a well-known function but it is not exploited today in image reconstruction algorithms.
- non-collinearity of the two gamma photons deriving from positron annihilation: momentum conservation implies that the two gamma rays resulting from the annihilation of a motionless positron are emitted on the same line of response

(LOR) in opposite directions. In practice the positron is not at rest when it is annihilated, which causes an average non-collinearity of the two gamma photons of about 0.25°. The error in the reconstruction of the emission point varies like the square of the scanner's radius. This error is reduced in equipment dedicated to the study of specific organs whose detectors can come as close as possible to the areas to be studied.

- size of the detection crystal (or pixel): it is the limiting factor in spatial resolution of commercially available scanners. Typically, the reconstruction error of each LOR is given by the half width of each pixel. The use of higher-density crystals and highly segmented photodetectors improves the spatial resolution. A significant increase in the number of channels, resulting from finer detector segmentation, implies that important efforts have to be made to develop cheap solutions for photodetectors and readout electronics. Difficult engineering problems have to be solved in order to integrate all the channels in a small volume and to keep the electronic equipment's thermal dissipation at an acceptable level.
- parallax effect: good spatial resolution has to be obtained not only on the axis, but also on the whole of the field of view (FOV). The depth of detecting crystals is limited by the density of the crystals and it cannot be reduced without altering the detector's sensitivity. If the conversion point of the gamma photon in the crystal is not known, spatial resolution deteriorates with increasing distance from the scanner axis. This error, which is known as parallax error, is increasing with decreasing the scanner's radius (Fig. 20.22). To limit this effect, one solution is to use several crystals in depth in a so-called phoswich configuration. If appropriate emission wavelength and decay time parameters are chosen for the crystals, the readout electronics can differentiate a conversion occurring in the front part or in the back part of the phoswich. Spatial resolution is therefore much more homogeneous on a larger field of view (Fig. 20.23). This scheme has been adopted in the ClearPET® small animal PET scanner with a combination of two 10 mm long LSO and LuAP crystals [31].

Another solution is to determine the γ conversion point in the crystal by means of a light sharing scheme with readout at both ends of the crystal. This solution has been chosen for 20 mm long LSO crystals in the ClearPEM®, a dedicated PET scanner for breast imaging [32].

20.3.2.3 Improving Time Resolution

Time-of-flight reconstruction can significantly reduce the signal to noise ratio in PET scanners, by constraining the annihilation point to a shorter segment on each LOR, with an uncertainty given by:

$$\Delta x = \frac{c}{2}\Delta t$$

where Δx is the position error, c is the speed of light, and Δt is the timing error.

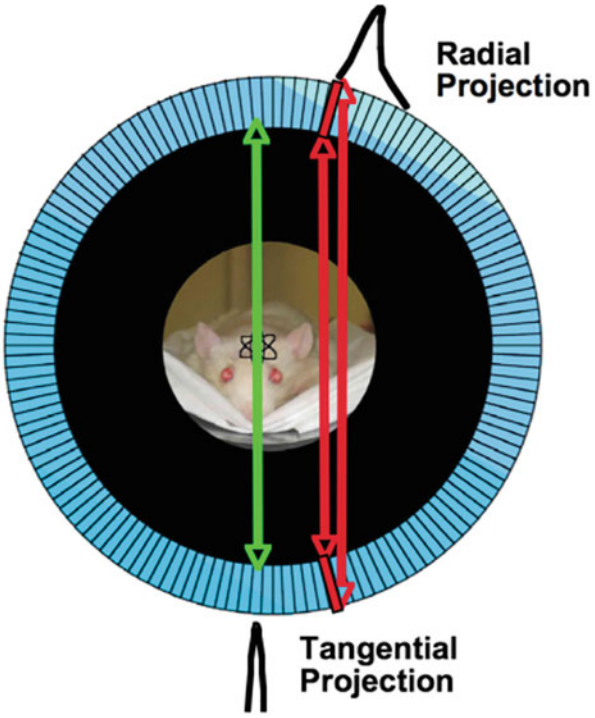


Fig. 20.22 Illustration of the parallax error for off centre events

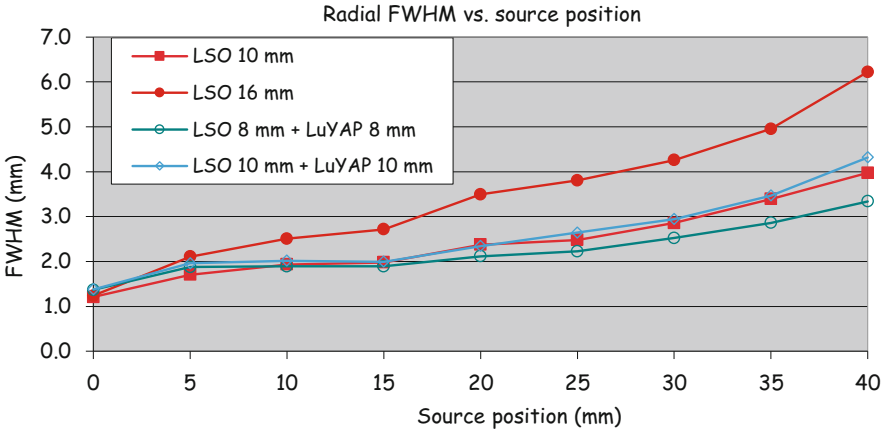


Fig. 20.23 Spatial resolution with and without phoswich for the ClearPET® (Courtesy Crystal Clear collaboration)

Until recently, PET scanners did not have any Time-of-Flight (TOF) capability to localize the position of the positron decay along the line of response (LOR) of the two γ-rays. Developments in fast scintillation crystals, photodetectors and electronics have open the way to TOFPET scanners with coincidence time resolution (CTR) improving progressively from 500–600 ps to 249 ps as recently announced by Siemens for their Biograph-Vision scanner. Pushing the limits of TOFPET techniques is motivated by the perspective for a significant improvement in the image signal-to-noise ratio (SNR), resulting in a corresponding clinical sensitivity increase and dose reduction, as given by the following equation:

$$SNR_{TOF} \Big/ SNR_{NONTOF} = \sqrt{\frac{2D}{c.CTR}}$$

where

D is the diameter of the Field of View (FOV),
c is the speed of light in vacuum
CTR is the Coincidence Time Resolution

Breaking significantly the 100 ps barrier, would dramatically improve the SNR (Fig. 20.24) and significantly remove artefacts affecting tomographic reconstruction in the case of partial angular coverage. This will open the field to a larger variety of organ-specific imaging devices as well as to imaging-assisted minimally invasive endoscopy.

Ultimately, a time resolution of 10 ps would lead to an uncertainty of only 1.5 mm for a given positron disintegration along the corresponding line of response (LOR). This is the order of accuracy achieved in today's very best small animal or organ

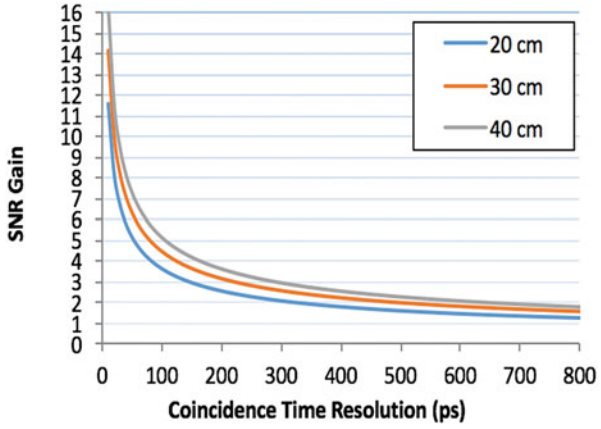


Fig. 20.24 Signal to Noise Ratio improvement as compared to non TOFPET as a function of the Coincidence Time Resolution for three different diameters of the Field of View (FOV)

specific PETs. The processing time of tomographic back-projection or iterative reconstruction algorithms would be considerably reduced, as true 3D information would be directly available for each decay event [33]. The possibility to see in real time the accumulation of the events during the acquisition could introduce a paradigm shift in routine clinical protocols, allowing in particular adapting the acquisition time to what is really observed and not to some predetermined evaluation. Moreover, such a timing resolution would allow recording the full sequence of all γ-ray interactions inside the scanner, including Compton interactions, like in a 3D movie, opening the way to the integration of at least a fraction of the Compton events in the image reconstruction, further improving the sensitivity.

To improve time resolution scintillating crystals with short decay time and fast treatment and acquisition electronics are needed. This has a double impact on image quality:

- as the width of the coincidence window is reduced, the number of isolated events decreases linearly. The proportion of random coincidences, which increase the detector's dead time and introduce noise into the image, is therefore reduced as the square of the single event rate. Images are less noisy and require less filtering, increasing spatial resolution and contrast.
- the use of time of flight information along the line of response (LOR) eliminates many random coincidences and reduces significantly the image noise.

In PET, the random event rate for an individual LOR is given by:

$$R = 2R_1R_2\Delta T$$

where R is the random event rate for that chord, R_1 and R_2 are the single event rates for two detector elements that form that chord, and ΔT is the width of the coincidence gate. The total number of random events in the image is the sum over all the chords, thus is proportional to ΔT. The mean contribution to the image from random events can be measured and subtracted, but the noise resulting from the statistical variations in this rate remains. The residual noise from random coincidences is usually estimated using the noise equivalent count rate (NECR) [34], a common figure of merit for comparing tomograph performance. The NECR is given by:

$$NECR = \frac{T^2}{T + S + 2R}$$

where *NECR* is the noise equivalent count rate, T is the true coincidence event rate, S is the scattered event rate, and R is random event rate. The noise equivalent count (NEC) metric is designed to obey counting statistics; that is, the NEC variance is equal to the NEC. Although the magnitude of the NECR is very sensitive to the source and camera geometries, this formalism is useful for predicting how changes in the trues, randoms, and scatters affect the image quality. Figure 20.25 shows such examples of NEC curves measured with a 20 cm long 20 cm diameter phantom

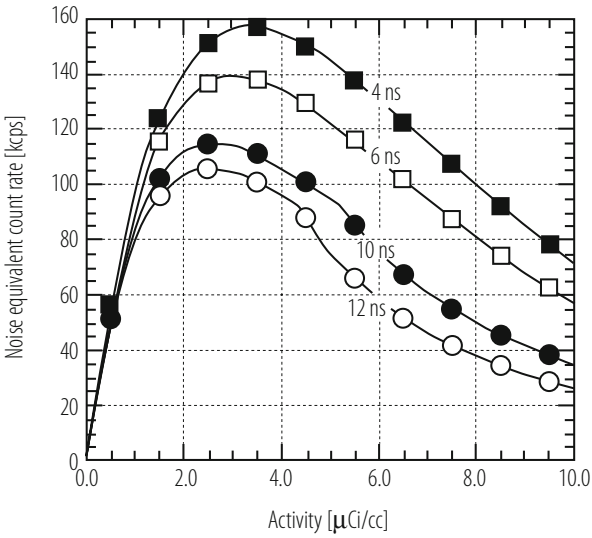


Fig. 20.25 NECR curves as a function of coincidence window width. The object imaged was a uniform 20 cm diameter cylinder and the camera had an 82 cm detector ring diameter and 15 cm axial extent (ECAT EXACT HR from Siemens) (Courtesy W.W. Moses, LBNL)

for a commercial PET scanner (ECAT EXAT HR from Siemens). The NEC value first increases linearly with the injected activity. It then progressively saturates and slowly decreases when the electronics dead time becomes significant as compared to the event rate. The importance of reducing the coincidence gate is evident from these plots.

20.3.3 Current and Future Technical Approaches

In the last few years, there have been noticeable improvements in commercial imaging equipment, with increased level of pixellization, better angular coverage, faster crystals, higher degree of integration of electronics with increased built-in functionality, more efficient reconstruction algorithms. Further progress is expected if medical imaging, and particularly nuclear imaging, can take advantage of significant technological advances in other fields like telecommunications or particle detectors.

Developments proceed along the following lines:

- new denser and faster scintillating crystals or direct conversion materials
- highly segmented and compact photodetectors
- low noise and highly integrated front end electronics
- data acquisition systems based on highly parallelized architectures
- efficient data filtering algorithms

- modern and modular simulation software based on universally recognized standards
- high performance image reconstruction and analysis algorithms

20.3.3.1 Conversion Materials and Metamaterials

The scintillating crystals used in PET scanners have to be dense, with a high atomic number, so as to optimize detection efficiency, and fast, in order to reduce dead time. The previous generation of PET scanners were using BGO crystal arrays, which have the advantage of being very dense (7.1 g/cm^3) and of having the highest atomic number known to this day for a scintillator (75), and therefore a high photoelectric conversion efficiency. Their main flaw is a slow decay time (300 ns) of the scintillating light. As a result, these scanners work with a limited sensitivity of about 1000 kcps/mCi/ml with a coincidence window of about 10–12 ns and a proportion of diffused events of more than 30%.

A new generation of scanners is now using LSO (Lutetium oxyorthosilicate) crystals [35], about 10 times faster than BGO and in some cases, the capability of determining the interaction depth in the crystals thanks to phoswich technology or double readout schemes. Combining these developments with progress in readout electronics and data acquisition, a gain in sensitivity by about one order of magnitude and in spatial resolution by a factor 2 or 3 has been achieved.

In the last 10 years, many groups, among them the Crystal Clear collaboration [36], have devoted a large effort on pluridisciplinary work to develop new scintillating materials meeting the demands for increasingly efficient detectors in physics and medical imaging. The most attractive scintillating crystals currently available or being developed for nuclear medicine are presented in Table 20.3. Cadmium Tungstate (CWO) and two ceramics compositions used for CT scanners are also mentioned for comparison. Figure 20.26 shows some pictures of the growth of LuAP ingots and pixel production developed in this context [37] for the preparation of phoswich pixels in combination with LSO for the ClearPET small animal PET scanner.

Other attractive crystals presently being developed are from the Lathanum halide group [38]. $LaBr_3$:Ce for instance has a higher light yield than CsI:Tl with more than 60,000 photons/MeV. The combination of high scintillation efficiency and good low energy linearity gives this crystal an unprecedented energy resolution (about 3% measured with avalanche photodiodes for 511 keV photons) and excellent timing properties.

Contrary to scintillators, semi-conductors convert the energy of the gamma photons to electric charge carriers (electrons and holes), which are directly collected on electrodes. However, most of the semi-conducting materials known today and used industrially, such as silicon, are not dense enough and do not have sufficient stopping power for 511 keV gammas (density 2.33 g/cm^3 and atomic number 14, as compared, for instance, to BGO density, 7.13 g/cm^3, and average atomic number,

Table 20.3 Scintillators already used or in development for medical imaging

Scintillator	Type	Density (g/cm^3)	Light yield (Ph/MeV)	Emission wavelength (nm)	Decay time (ns)	Hygroscopic
NaI:Tl	Crystal	3.67	38,000	415	230	Yes
CsI:Tl	Crystal	4.51	**54,000**	550	1000	Lightly
CWO	Crystal	7.9	28,000	470/540	20,000/5000	No
$(Y,Gd)_2O_3$:Eu	Ceramics	5.9	19,000	610	1000	No
Gd_2O_2S:Pr,Ce,F	Céramics	**7.34**	21,000	520	3000	No
BGO	Crystal	**7.13**	9000	480	300	No
GSO:Ce	Crystal	6.7	12,500	440	60	No
LSO:Ce	Crystal	**7.4**	**27,000**	420	**40**	No
LuAP:Ce	Crystal	**8.34**	10,000	365	**17**	No
$LaBr_3$:Ce	Crystal	5.29	**61,000**	358	35	Very

Particularly attractive parameters are marked in bold

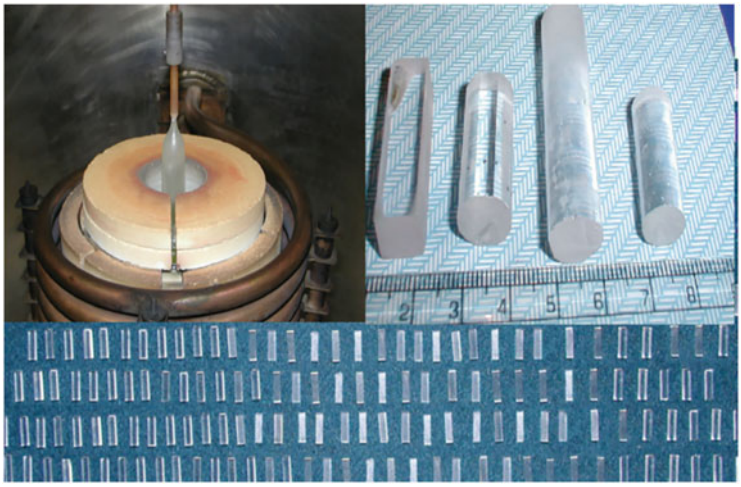


Fig. 20.26 LuYAP crystals produced in Bogoroditsk, Russia (Courtesy Crystal Clear, CERN)

75). This technique is nevertheless used in single photon X-ray imaging and makes the acquisition of high resolution and high contrast digital images possible [39].

For gamma imaging, multi-layer systems could be considered, but to this day integrating a huge number of channels in these conditions has not been solved, especially in terms of connectivity. Yet interesting solutions to these problems have recently become available through recent developments on pixel detectors for tracking devices. For example, using bump-bonding techniques semiconductors are coupled directly to their readout electronics. It has also become possible to integrate the semiconductor directly with ASIC readout chips and to read a large number of channels on a very small surface quickly and with low noise.

New semiconducting materials, denser than silicon, are also being developed: Gallium Arsenide (GaAs) [40], with a density of 5.32 g/cm^3 and an average atomic number of 31, Cadmium Telluride (CdTe), with a density of 5.85 g/cm^3 or Cadmium and Zinc Telluride (CdZnTe, or CZT) [41], with a density of 5.78 g/cm^3 but whose atomic number is higher, 49 instead of 32. One of these materials is particularly attractive because of its density and high atomic number: Mercuric Iodide (HgI_2). With a density of 6.4 g/cm^3 and an average atomic number of 62, it nearly equals the stopping power of the best scintillating crystals (BGO, LSO and LuAP). It is unfortunately very difficult to grow in reasonable size and consistent quality.

One remaining problem for these materials is the limited charge collection speed and efficiency, which requires well designed geometries with small drift regions and optimized, cost effective production technologies with a very good control of charge carriers traps.

As shown in [33], the ambitious target of a few tens of picoseconds Time-of-Flight resolution can only be met with scintillators exhibiting a very fast rise time in the scintillation process and the possibility to combine standard scintillation processes with a few hundreds of prompt photons generated by another mechanism. One of the very attractive mechanisms for the production of sub-ns scintillation processes is related to quantum confinement in nanocrystals, as explained in [42, 43].

The challenge is to optimize the design of a metamaterial combining the high density and stopping power (small radiation and interaction lengths, small Moliere radius, high photoelectric fraction) of already well known scintillators (LSO, L(Y,G)SO, PWO, BGO, LuAG, YAG, GGAG, Lu(Y)AP, etc . . .) to the ultrafast (<1 ns) light emission of nanocrystals. Different solutions are presently under study for an optimal combination of these two classes of materials, solving at the same time the problem of light transport by the use of photonic fibers, as proposed in [44].

20.3.3.2 Photodetectors

In nuclear medicine the basic technique to detect ionizing radiation uses scintillators to convert X- or gamma-rays into light and then into an electric signal by a photodetector. Until recently, the standard commercial imaging cameras were equipped with photomultiplier tubes (PMT) used as light sensors. However, these technologically mature products are approaching their limits in terms of dimension, efficiency and cost. However, their sensitivity to magnetic fields prevents their use in combined PET/MRI devices. The trend toward larger numbers of scintillating pixels of increasingly smaller size will limit their use in the future.

New compact photodetectors have been developed over the last two decades, for instance hybrid photodetectors (HPD), photodiodes and avalanche photodiodes, thanks to which sensitivity, spatial resolution and immunity to magnetic field could be significantly improved. Arrays of avalanche photodiodes have been considered

for hybrid PET/MRI scanners and several prototypes dedicated to brain and breast imaging have been built.

Avalanche photodiodes suffer however, a number of drawbacks, such as a limited gain of a few hundred, a large excess noise factor and relatively poor timing characteristics preventing their use in PET Time-of-Flight systems. For these reasons the intense ongoing R&D activity on multipixel Geiger mode avalanche photodiodes (also called Silicon Photomultipliers or SiPM) is followed with particular attention. Their working principle is based on the segmentation of the large coupling area with the scintillating crystal into a large number of small avalanche photodiode cells working in Geiger mode and connected in parallel via individual quenching resistors. The first devices of this type were developed in the late 1990s in Russia and since then several designs have been realized [44, 45].

The cells in a SiPM are all identical with dimensions ranging from 7×7 to $70 \times 70\ \mu m^2$. Each cell operates as an independent photon counter in Geiger-mode when the bias voltage is 10–20% higher than the breakdown voltage and behaves as a binary device since the signal from a cell always has the same shape and amplitude.

The gain is similar to the one of a photomultiplier, in the range of 10^5 to 10^7. Since each cell acts as a digital single photon counter the excess noise factor is very small. The light yield is directly given by the number of fired cells. This assumption is of course valid only, if crosstalk between individual cells can be eliminated, which has been solved by the use of trenches between the pixels.

Present SiPMs have a dead-time per cell of the order of several μs. For the device to be linear to the light response of fast scintillators having a decay time in the ns range, the number of cells must be larger than the maximum number of photons per event. This requires SiPM to have a cell density above 1000 cells/mm^2. SiPMs with 100–10,000 cells/mm^2 are currently available.

The overall efficiency of the device depends on the quantum efficiency of each individual cell, which is wavelength dependent but is now reaching 60–65% at the emission wavelength of LSO, and with the geometrical acceptance due to the dead space between the cells, which ranges between 20% and 70% depending on the design.

One of the most promising features of SiPMs for medical imaging applications is related to their excellent timing resolution. The active layer of silicon in a SiPM is very thin (2–4 μm), the avalanche breakdown process is fast and the signal amplitude is large. Impressive timing resolution of about 10 ps for single photons have been reported. For a system of a SiPM and a LYSO crystal with dimensions of $2 \times 2 \times 3$ mm^3, $2 \times 2 \times 10$ mm^3 and $2 \times 2 \times 20$ mm^3, a time resolution of 73 ps, 100 ps and 122 ps FWHM respectively has been measured for 511 keV X-rays [46].

20.3.3.3 Highly Integrated Low Noise Front-End Electronics

The large number of readout channels requires highly integrated low-noise and high-speed readout electronics, typically using integrated circuits of the VLSI CMOS type [47]. Institutes of particles physics are experienced in designing and integrating

large numbers of multichannel and multifunction low noise and fast electronics into complex detector systems.

Medical imaging should also benefit from the progress in large-scale integration of electronic channels with complex functions and highly segmented sensors. The concept of a hybrid detector, in which each pixel is integrated directly to its readout electronics opens totally new perspectives in the conception and architecture of new imaging systems.

20.3.3.4 Highly Parallelized and Intelligent Data Acquisition System (DAQ)

The data acquisition system has a double function: first, it has to discriminate between the interesting events—coming from real X- or γ-ray interactions—and the various types of background, and, secondly, it has to transfer data to the computer(s) where these events will be processed. In medical imaging, most of the "real events" are triggered by the time and amplitude analysis of each sensor. In this acquisition system, data are selected, standardized, organized, corrected, processed with more or less complex algorithms and finally presented as an image file.

In data acquisition system with conventional architecture, data are treated sequentially. A new event is only accepted, once the processing of the previous event is completed; pile-up is thus avoided, but dead time occurs at each treatment stage affecting the data collection efficiency and coming from three main sources: first, the sensor and the electronic pulse generation system, second, the analog-to-digital conversion of this signal, and third, the logic treatment (in general, the major one).

Although pipeline architectures could be seen as more complex and more expensive than conventional ones, improvements in electronics (ASIC and FPGA) in terms of integration and cost suggest that in a very close future medical imaging devices entirely based on this concept might be designed. Similarly, progress in data transfer speed between the electronic system and the analysis system is no more a limiting factor for data transfer at the 1 Gbit/s level. Finally, with the parallelization of processors in cheap PC clusters (processor farms), adequate processing power is now available.

In the future, data acquisition system will no longer be a limiting factor in medical imaging.

20.3.3.5 Simulation Software

Monte Carlo simulation methods are an essential tool for developing new detectors in medical imaging.

Versatile generic simulation tools have been developed for particles physics, for instance Geant4 and FLUKA at CERN and INFN, EGS4 at SLAC or MCNP at Los Alamos National Laboratory. More recently, the development of Geant4

has made it possible to include efficient geometrical modelling and visualisation tools. GATE [48] was developed on this basis to simulate PET and SPECT imaging devices; it is a simulation platform written specifically to model imaging systems, through which time-dependent phenomena—detector motion, decrease in isotope radioactivity, dead time phenomena—can be followed. This new simulation tool, developed, validated and documented by the OpenGATE collaboration regrouping about 20 laboratories of medical and particles physics, is freely available on the Internet and is currently used by a community of more than 200 scientists and industry worldwide.

20.3.4 Image Reconstruction Algorithms

The data provided by transmission and emission tomography make it possible to reconstruct projections of the image, which are then combined to images thanks to tomographic reconstruction methods [49]. There are two possible approaches to deal with tomographic reconstruction problems. The first one is analytical and consists in treating the measured projections as if they were perfect mathematical projections. In this case, it is necessary to make a number of hypotheses on linearity and continuity, but the data, often incomplete and noisy, will not always fulfil these requirements and give rise to artefacts. The second approach is phenomenological. It consists in modelling the measuring process using a probability matrix, which has to be inverted through iterative algebraic techniques. Such algorithms iterate a process aiming at optimising an objective function, for instance the verisimilitude function coming from the Poissonian nature of the data recorded by the tomograph. This is the case for instance of the maximum likelihood expectation maximization (MLEM) algorithm and its variant using ordered subsets expectation maximization (OSEM). These iteration methods are less demanding on the geometry of the detector, and do not require a complete set of projections. On the other hand, they require high calculation power. Fortunately, in the near future the Grid or Cloud will provide considerable and massively distributed computing resources. The quality of the results of an iterative process are more difficult to evaluate than the results of an analytical one. Monte Carlo simulation is often a key element to study and optimise these algorithms.

20.4 Multimodality

20.4.1 Need for a Multimodal Approach

SPECT or PET scanners allow localizing radiotracers uptakes in the human body and are, as such, very powerful tools for basic research in cognitive sciences, for

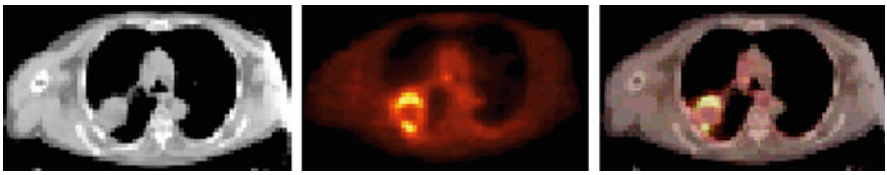


Fig. 20.27 Primary lung cancer imaged with a PET/CT scanner. A large lung tumour, which appears on CT as a uniformly attenuating hypo dense mass (left), has a rim of FDG activity and a necrotic centre revealed by PET (middle). The combined image (right) allows a precise localization of the active parts of the tumour (Courtesy Dave Townsend, University of Tennessee)

clinical oncology and cardiology and for kinetic pharmaceutical studies. However, they do not deliver precise anatomical images, like MRI or X-ray CT for instance. Whilst effective software image fusion techniques exist there is a great deal of interest in performing functional and anatomical studies as simultaneously as possible in order to improve registration accuracy and to resolve the logistical problems associated with software registration.

Modern scanners combine the very high sensitivity of PET for metabolic imaging with the high spatial resolution anatomic information delivered by X-ray CT or another anatomical modality. Indeed, the majority of PET and SPECT systems currently being installed now incorporate a CT scanner in the same gantry, so that functional and anatomical images can be performed in rapid succession. Features identified with PET or SPECT can then be accurately localized via the CT scan. The CT data can also be used to determine the photon attenuation correction, an advantage, in particular for overweight patients. These machines provide impressive images giving the very precise localization of the metabolic activity of organs and tumours (Fig. 20.27).

The development of bimodal acquisition systems, for metabolic, functional and anatomical data, like PET/CT combined scanners [27]—several thousand machines exist as of today—, is radically modifying the patients care thanks to increased precision in diagnosis. This approach also decreases significantly the number of imaging scans for a patient.

Combined PET/CT imaging brings additional benefits in the planning of radiotherapy, which is a promising area for research and clinical applications. The principle of radiation therapies is to modulate the intensity according to the spatial distribution of the area to be treated (Intensity-Modulated Radiation Therapy, IMRT). The combination of PET images, which provide information on the metabolic extension and heterogeneity of tumoural tissues, and CT images, which provide precise structural information and location of the tumour, helps defining an irradiation map to focus the therapy on the particularly active areas of the tumour.

Multimodality has become an intensive research area, the challenge being to take the best advantage in the combination of anatomic and functional information by optimizing the choice of the imaging modalities as a function of the application domain. The progress in SiPM technology has opened the way for a vibrant

field of development for PET/MRI systems, in particular for brain studies, taking advantage of the high functional sensitivity of the PET and of the very good soft tissue contrast capability of the MRI. Similarly, coupling optical fluorescence methods and PET or SPECT is very attractive for studying biologic processes on small animals. In another domain a Cerimed [50] collaboration has designed a PET/SPECT/Ultrasound dedicated breast imaging camera allowing to simultaneously access a variety of parameters on breast tumours such as energetic metabolism of cells, response to specific hormonal ligands (herceptin, bombesin), echogenecity, elastometry, Doppler, and to correlate them in order to optimize the treatment plan of the patients.

Finally, the combination of PET and SPECT imaging associated to the labeling of various ligands signaling different molecular pathways opens the way to multifunctional imaging. Such an approach could prove useful for identifying the molecular profiling of a tumour in a single exam. It would also allow the simultaneous recording of the expression of several neurotransmitters under specific stimuli, a very powerful approach in cognitive sciences.

20.4.2 Outlook: Towards Integrated Morphologic and Functional Imaging

Biological systems are so complex that there is an important need to develop imaging modalities capable of simultaneously recording different molecular pathways in a quantitative and dynamic manner. Helping to address this issue, multi-parametric molecular imaging involves combining the excellent sensitivity and specificity of molecular imaging (PET or SPECT) with a complementary high-spatial resolution imaging modality (CT, MRI or ultrasound).

The most frequently used equipment combines a PET scanner and an X-ray CT. At present, these combined systems are large, consisting of independent scanners mounted in-line in a common gantry, not generally mounted on the same rotary holder. This results in some imprecision in the image fusion process due to external and internal movements of the patient. CT data provide crucial information for the correction of the unavoidable attenuation factors from the patient's body in PET images and for improving image quality by decreasing the influence of artefacts. Partial volume effects are caused by PET's limited spatial resolution, which dilutes information from small hot spots onto a larger area because of the blurring of the image. CT information, which provides much more precise anatomical information, helps to correct, at least partially, these negative effects. This correction is difficult if both imaging systems acquire data in distinct, poorly correlated spaces. In the case of attenuation and partial volume correction, it is crucial to record both data sets as simultaneously as possible so as to guarantee good image superposition. A major challenge is to further integrate the readout of X-rays and γ-rays. Simultaneous recording of anatomical (CT) and functional (PET and/or SPECT) information by

the same reading head is in principle possible thanks to progress in microelectronics. The large functionality of modern ASIC's makes it possible to develop electronic readout channels able to count each individual event, well suited for CT, PET and SPECT signal treatment. This is a particularly interesting perspective, because it would make it possible to correct attenuation and partial volume parameters more precisely.

Another way of obtaining images associating anatomical and functional information is to merge MRI and PET images. Again both data sets have to be acquired as simultaneously as possible, even if a universal acquisition system cannot be considered here because of the fundamental differences between these two modalities. The PET/MRI approach is particularly promising for brain studies. Indeed, MRI gives much better images than CT for the soft brain tissues behind the skull.

Besides, BOLD contrast MRI, which relies on the variation of blood oxygenation level, is promising as a tracer of neuronal activity in functional MRI imaging. Combining this approach with PET functional imaging using various ligands (dopamine, serotonine, acetylcholine, glutamates, opiates, etc.) opens the way to a better understanding of fundamental neurotransmission mechanisms in the brain.

However, a number of significant technological problems arise from recording almost simultaneously MRI and PET images; these problems are mainly caused by the presence of powerful magnetic fields in MRI, with a high homogeneity requirement. To combine the two systems, innovative technologies are needed. The development of SiPM matrices has solved the problem of photodetectors having to be immune to the magnetic field. Moreover, they are extremely compact and require an operating voltage of only a few hundred volts. However, a number of other constraints remain: in a PET/MRI system conducting or ferromagnetic materials must also be carefully avoided because they would alter the homogeneity of the MRI magnetic field. Other technical difficulties, which have to be solved are linked to gradient coils and to MRI's radiofrequency fields, which require effective shielding for PET parts against Eddie currents and electromagnetic noise.

Thanks to multiparametric molecular imaging, a radical shift is currently taking place in the way diseases are managed: from the present onefitsall approach to one that delivers medical care tailored to the needs of individual patients. This includes the detection of disease predisposition, early diagnosis, prognosis assessment, measurement of drug efficacy and disease monitoring. Thus, the introduction of personalized medicine requires an unprecedented effort to develop new technologies in fields of diagnostic and image-guided therapeutic medicine (theranostics) including pathology and imaging. Such imaging tools should characterize diseases and assess treatment efficacy, with the added advantage of non-invasive monitoring at multiple time points. The recent explosion of molecular biology and imaging technologies is now allowing simultaneous quantitative and dynamic characterization of several biological processes inside the body at the molecular and genetic level. This exciting new field will transform the future of medicine on a massive scale and will have an enormous impact on the advancement of targeted therapies for personalized medicine.

References

1. C.C. Ling, J. L. Humm, S. M. Larson, H. Amols, Z. Fuks, S. Leibel, J. A. Koutcher, Towards multidimensional radiotherapy (MDCRT): biological imaging and biological conformality. *Int. J. Radiat. Oncol. Biol. Phys.* 2000; 47:551–560.
2. R. Orecchia, A. Zurlo, A. Loasses, M. Krengli, G. Tosi, P. Zucali, S. Zurrida, and U. Veronesi, *Particle Beam Therapy (Hadrontherapy): Basis for Interest and Clinical Experience*, European Journal of Cancer 34: 456 –468 (1998).
3. W. Enghardt, P. Crespo, F. Fiedler, R. Hinz, K. Parodi, J. Pawelke, F. P.onisch, Charged hadron tumour therapy monitoring by means of PET, NIMA 525, 284-288 (2004).
4. Deych D, Dobbs J, Marcovici S, Tuval B (1996), Cadmium tungstate detector for computed temography. In: Dorenbos P, van Eijk CWE (eds). Inorganic scintillators and their application. Delft University Press, pp 36–39.
5. C. Greskovich, S. Duclos, Annu. Rev. Mater. Sci., 27(1997)69.
6. Kostler W, Winnacker A, Rossner W, Grabmaier BC (1993), Effect of Pr-codoping on the X-ray induced afterglow of $(Y,Gd)_2O_3$:Eu, J Phys Chem Solids 56: 907–913.
7. C. Greskovich, D. Cusano, D. Hoffman, R. Riedner, American Ceramic Society Bull., 71(1992)1120.
8. E. Gorokhova, V. Demidenko, O. Khristich, S. Mikhrin, P. Rodnyi, J. Opt. Technology, 70(2003)693.
9. Y. Ji, J. Shi, J. Mater. Res., 20(2005)567.
10. S. Sekine, T. Yanada, U.S. Patent No. 6,876,086 B2, 2005.
11. R. Luhta, R. Mattson, N. Taneja, P. Bui, R. Vasbo, in: Medical Imaging 2003: Physics of Medical Imaging, Proc. Of SPIE 5030(2003)235.
12. A. Goushcha, A. Popp, E. Bartley, R. Metzler, C. Hicks, in: Medical Imaging 2004: Physics of Medical Imaging, Proc. Of SPIE 5368(2004)586.
13. F. Natterer, The Mathematics of Computerized Tomography, New York: Wiley, 1986.
14. A.C. Kak, M. Slaney, Principles of Computerized Tomographic Imaging, New York: IEEE Press, 1988.
15. Guang-Hong Chen et al., A novel extension of the parallel-beam projection-slice theorem to divergent fan-beam and cone-beam projections, Med. Phys. 32 (3), March 2005.
16. Crawford, C. R., King, K. F.: Computed tomography scanning with simultaneous patient translation. Med. Phys. 1990; 17:967-82.
17. T. G. Flohr, K Stierstorfer, S. Ulzheimer, H. Bruder, A. N. Primak, C. H. McCollough, Med. Phys., 32 (2005) 2536.
18. N.J. Pelc, Recent and future directions in CT imaging, Ann. Biomed Eng. 2014 Feb: 42(2): 260-268.
19. D. Fornel, Technology Improvements in Current Generation CT systems, ITN September 2015, https://www.itnonline.com/article/technology-improvements-current-generation-ct-systems.
20. Wieczorek H., Frings G., Quadfield P., et al.(1995) CsI:Tl for solid state X-ray detectors, Proc. In Dorenbos P, van Eijk CWE (eds). Inorganic Scintillators and Their Applications, Delft University Press, 547-554.
21. X. Llopart et al., Timepix, a 65k programmable pixel readout chip for arrival time, energy and/or photon counting measurements, Nucl. Intr. And Meth. A (2007), doi:https://doi.org/10.1016/j.nima.2007.06.054.
22. M. Titov, New developments and future perspectives of gaseous detectors, Nucl. Intr. And Meth. A (2007), doi:https://doi.org/10.1016/j.nima.2007.07.022.
23. R.K. Swank, Absorption and noise in X-ray phosphors, J. Appl. Phys. 44, 4199-4203 (1973).
24. M.G. Bisogni et al., NIMA 546, 14 (2005).
25. P. Delpierre, A history of hybrid pixel detectors, from high energy physics to medical imaging JINST 9 C05059, 2014.
26. F. Cassol et al., Characterization of the imaging performance of a micro-CT system based on the photon counting XPAD3/Si hybrid pixel detectors, Biomed. Phys. Eng. Express 2 (2016) 025003.

27. H. Schöder et al., PET/CT: a new imaging technology in nuclear medicine, Eur J Nucl Med Mol Imaging (2003) 30:1419-1437.
28. T. Jones, Molecular imaging with PET – the future challenges, The British Journal of Radiology, 75 (2002), S6–S15.
29. European Society of Radiology (ESR), Medical imaging in personalised medicine: a white paper of he research committee of the European Society of Radiology (ESR), Insight Imaging, Apr. 2015, 6(2): 141-155.
30. T. Jones, D. Townsend, History and future technical innovation in positron emission tomography, J. Med. Imag. 4(1), 011013 (2017).
31. P. Sempere Roldan et al., Performance Evaluation of Raytest ClearPET®, a PET Scanner for Small and Medium Size Animals, Conference records of the IEEE NSS/MIC conference, Oct. 27-Nov. 3, 2007, Hawaï, 2859-2864.
32. M. C. Abreu et al., Clear-PEM: A PET imaging system dedicated to breast cancer diagnostics, NIMA 571, 81-84 (2007).
33. P. Lecoq, Pushing the limits in Time-of-Flight PET imaging, IEEE Transactions on Radiation and Plasma Medical Sciences, Vol. 1, N°6, November 2017.
34. S. C. Strother, M. E. Casey, and E. J. Hoffman, "Measuring PET scanner sensitivity: Relating count rates to image signal-to-noise ratios using noise equivalent counts," *IEEE Trans. Nucl. Sci.*, vol. NS-37, pp.783-788, Apr. 1990.
35. C.L. Melcher et al., A promising new scintillator: cerium doped lutetium oxyorthosilicate, Nuclear Instruments and Methods in Physics Research A 314 (1992) 212-214.
36. Crystal Clear Collaboration, RD18, CERN/DRDC/P27/91-15.
37. C. Kuntner et al., Advances in the scintillation performance of LuYAP:Ce single crystals, Proceedings of the 7th conference on Inorganic Scintillators and their Use in Scientific and Industrial Applications, Valencia, Spain, Sept 2003, Nuclear Instruments and Methods in Physics Research A 537 (2005) 295-301.
38. V.D. van Loef et al., Scintillation properties of $K_2LaX_5:Ce^{3+}$ (X=Cl, Br, I), Proceedings of the 7th conference on Inorganic Scintillators and their Use in Scientific and Industrial Applications, Valencia, Spain, Sept 2003, Nuclear Instruments and Methods in Physics Research A 537 (2005) 232-236.
39. JC Clemens et al., PIXSCAN: CT-Scanner for Small Animal Imaging Based on Hybrid Pixel Detectors. To be published in conf rec 7th International Workshop on Radiation Imaging Detectors, IWORID-7, July 4-7, 2005 in Grenoble, France.
40. J. C. Bourgoin, A new GaAs material for X-ray imaging,. Nuclear Instruments and Methods in Physics Research A460 (2001), 159-164.
41. A. Owens et al., The X-ray response of CdZnTe, Nuclear Instruments and Methods in Physics Research A484 (2002), 242-250.
42. R.M. Turtos et al., "Timing performance of ZnO:Ga nanopowder composite scintillators", Phys. Status Solidi RRL 10, No. 11, 843–847 (2016).
43. R.M. Turtos et al., "Ultrafast emission from colloïdal nanocrystals under pulsed X-ray excitation", *JINST_068P_06*.
44. P. Lecoq, "Metamaterials for novel X- or γ -ray detector designs," in Proc. IEEE Nucl. Sci. Symp. Conf. Rec., Dresden, Germany, 2008, pp. 680–684.
45. I. Britvitch et al., Development of scintillation detectors based on avalanche microchannel photodiodes, Proceedings of the 1st international conference on Molecular Imaging Technology, Marseilles, France, May 9-12, 2006, Nuclear Instruments and Methods in Physics Research A 571 (2007) 317-320.
46. S. Gundacker et al., "State of the art timing in TOF-PET detectors with LuAG, GAGG and L(Y)SO scintillators of various sizes coupled to FBK-SiPMs", *2016 JINST 11 P08008.*
47. E.H.M. Heijne, Future semiconductor detectors using advanced microelectronics with post-processing, hybridization and packaging technology, Nuclear Instruments and Methods in Physics Research A 541 (2005) 274-285.

48. D. Strul et al., “GATE (Geant4 Application for Tomographic Emission): a PET/SPECT general-purpose simulation platform,” Nucl. Phys. B (Proc. Suppl.) 125C (2003) 75-79. (http://www.opengatecollaboration.org)
49. B. Bendriem and D.W. Townsend, The theory and Practice of 3D PET, Kluwer Academic Publishers, 1998, ISBN 0-7923-5108-8.
50. Cerimed, European Centre for Research in Medical Imaging, based in Marseille, France. http://www.cern.ch/cerimed/.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 21
Solid State Detectors for High Radiation Environments

Gregor Kramberger

21.1 Introduction

The solid state particle detectors emerged in 1950 [1]. Initially Si and Ge detectors operated as junction diodes were used for charged particle detection and γ spectroscopy measurements (Chap. 5). Although these detectors are superior to gaseous detectors in many respects, being a crystalline medium meant that they are susceptible to radiation damage. Unlike in gaseous detectors where the detection media can be exchanged the semiconductor crystals have to retain their detection properties over the entire envisaged period of operation. The particle detection capabilities and the energy resolution degrade gradually with irradiation, which limits their lifetime.

A large majority of present high energy experiments uses position sensitive silicon detectors which became widely available after the introduction of planar process in 1980 [2]. Their goal is achieving desired position resolution with as few read out channels as possible, while keeping detection efficiency close to 100%. At the present and particularly future experiments high particle rates close to the interaction point require very fine segmentation and high position resolution of detectors in order to be able to associate hits with tracks.

In the future a precise timing information associated with a track and even with each sensor hit may be required to cope with large multiplicity of tracks. The sensor hits and associated tracks will be therefore separated not only spatially, but also in time allowing easier assignation of tracks to different collisions occurring within each colliding particles bunch crossing.

G. Kramberger (✉)
Experimental Particle Physics, Jožef Stefan Institute, Ljubljana, Slovenia
e-mail: gregor.kramberger@ijs.si

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_21

High particle rates cause radiation effects. The most important is the damage of the crystal lattice which leads to the degradation of the measured charge after passage of ionizing radiation. At the same time the noise may increase for various reasons thus significantly reducing the signal-to-noise ratio. Consequently the detection efficiency, energy, and position resolution may degrade to the level where the detectors become unusable. Extensive research was made in the last decades to understand the damage in silicon detectors and to manipulate the properties of silicon aiming at radiation-harder detectors. The research was not only limited to silicon but alternative semiconductor materials were considered.

It is not only the bulk crystal that is affected by irradiation, but also the surface. The radiation effects at the silicon—silicon oxide interface not only change the performance of silicon detectors, but are the main reason for radiation damage of electronics. The latter was often considered a bigger problem than the radiation damage of detectors, particularly in environments where the ionization dose was large (e.g. synchrotron radiation). With the advent of deep sub-micron CMOS processes, electronics was thought to became intrinsically radiation hard and no special radiation hardening processes would be required. An important contribution to the radiation hard electronics was also introduction of radiation-tolerate design rules. However, for very small feature sizes, e.g. very deep sub-micron processes, such as 0.130, 0.065 μm, radiation hardness of electronics, rather than sensors, could become a limiting factor at harshest radiation environments.

On the other hand the effects of radiation damage were exploited for dose measurements. Active dosimeters appeared for both measurements of ionizing and non-ionizing energy losses in silicon crystal such as $p-i-n$ diodes [3] and radiation sensitive field effect transistors [4].

21.2 High Radiation Environments

The radiation environments differ in composition and energies of the particles producing the radiation damage. Although the particles that are to be detected contribute largely to the damage it is often the background particles that dominate. As will be described later the damage depends on the type of the particle. While X-rays alter the properties of the detector surface they can not displace the semiconductor atoms from the lattice. On the other hand neutron irradiation affects only the lattice and energetic charged hadrons and leptons damage both the lattice and the surface. The difference in damage creation and its effects to detector operation will be discussed later. First we review the radiation environments where particle detectors are employed.

Collider Experiments In general there are three major types of accelerators with respect to collision particles: hadron ($p-p$, $\bar{p}-p$, heavy ions), lepton (e^+-e^-, eventually $\mu^+-\mu^-$) and lepton-hadron ($e^{+,-}-p$). The flux of particles traversing

the detectors is given by the particles originating from the collisions (ϕ_{coll}) and secondary radiation that originates from the spectrometer or the accelerator (ϕ_{sec})

$$\phi = \phi_{coll} + \phi_{sec}, \tag{21.1}$$

The flux of particles crossing the detectors is much larger at hadron colliders than at lepton colliders, owing to a difference in total cross-section σ_{tot} of colliding particles. The radiation environment at lepton colliders is dominated by $e^{\pm}$ from Bhaba scattering. Consequently the radiation damage of detectors at hadron colliders is much more severe than at lepton colliders.

A significant secondary irradiation, particularly at high luminosity colliders, can arise from back-scattered neutrons originating in breakup of nuclei in calorimeters and other parts of spectrometers after interaction with highly energetic hadrons. The secondary radiation originating from the accelerator such as synchrotron radiation, beam-gas interactions or halo particles scraping the collimators should be small but can represent in case of an accident a significant contribution to the total fluence Φ (integral of flux $\Phi = \int \phi \, dt$) of particles traversing the detectors.

The required radiation tolerance/hardness of vertex detectors at different colliders is given in the Table 21.1. Placing of the detectors in the spectrometer determines their exposure. The ϕ_{coll} decreases quadratically with the distance from the interaction point. The large cross-section for soft collisions result in larger ϕ_{coll} at small angles with respect to beam. Large ϕ_{coll} at small angles is also characteristic for asymmetrical beams (energy, particle) or fixed target experiments. A particle fluence profile for ATLAS experiment [5] at the Large Hadron Collider (LHC) is shown in Fig. 21.1. The dominating particles are at small radii mainly pions and protons originating from collisions and "albedo" neutrons from the calorimeters for $R > 20\,\text{cm}$.

Table 21.1 The review of basic parameters of some accelerators and required radiation hardness of the most exposed detectors for the entire operation period

Accelerator	Type	σ_{tot} [barn]	$\mathcal{L}$ [$\text{cm}^{-2}\,\text{s}^{-1}$]	$\sim \int \phi dt$ [$\text{n}_{eq}\,\text{cm}^{-2}$]	Dose in Si [Gy]
Super KEK-B	e^+-e^- (8,3.5 GeV)	4n	$5.0 \cdot 10^{35}$	$<2 \cdot 10^{12}\,\text{cm}^{-2}$	<10 k
ILC	e^+-e^- (250,250 GeV)	3p 3p	few 10^{34}	$\sim 10^{10}$	few k
HERA	$e^{+,-}$-p (27.5, 920 GeV)	10^{-3} ($Q^2 < 100\,\text{GeV}$)	$7 \cdot 10^{31}$	$<10^{13}$	<2 k
Tevatron	$\bar{p}$-p (0.98,0.98 TeV)	70 m	$1.7 \cdot 10^{32}$	$<10^{13}$	<30 k
LHC HL-LHC (>2026)	p-p (7,7 TeV)	100 m	$10^{33} - 10^{34}$ $5 - 7.5 \cdot 10^{34}$	up to $5 \cdot 10^{15}$ up to $2 \cdot 10^{16}$	~2.5 M ~10 M
FCC	p-p (50, 50 TeV) foreseen >2040	100 m	$5 - 30 \cdot 10^{34}$	up to $6 \cdot 10^{17}$	~400 M

The total cross-section without Bhaba scattering is given for $e^+ - e^-$ accelerators

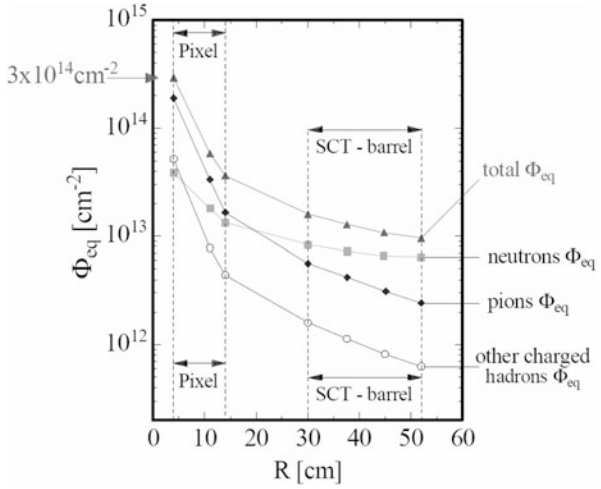


Fig. 21.1 Yearly fluence profile in ATLAS experiment at LHC design luminosity. The radiation damage caused by different particles was used to normalize the fluences (see next section for Φ_{eq}). The arrows denote the location of the pixel and strip detectors (SCT)

The choice of detector technology at a given radius depends on the ability to retain the detection efficiency and position resolution at required levels. At the same time the material budget should be kept low in order not to spoil the tracking performance. At many experiments the most exposed detectors are beam position/condition and radiation monitors (Chap. 18).

Space Applications

Particle detectors are an important constituent of many space missions. They are mainly used as spectrometers, visible light detectors and charged particle trackers. The radiation fields are far less severe than that at accelerators experiments, but the detectors and the information that they provide can be far more susceptible to the radiation effects (e.g. CCD, DEPFet, Si-drift detectors). The origin of radiation in space comes from three sources:

- **Galactic cosmic radiation;** Consists primarily of nuclei (85% protons, 14% Helium, 1% heavier ions among which Fe and C are most abundant ones). The relevant particles for damage creation have energies between 1–20 GeV. The fluxes of cosmic particles are shown in Fig 21.2a. The flux depends on the activity of the sun through interaction with solar wind (a continuous stream of high ionized plasma emerging from the sun). Interactions of highly energetic particles with nuclei in the earth's atmosphere or space-vessel produce showers of ionizing particles which increase the intensity of the radiation.
- **Solar particles;** The sun is also a sporadic source of lower energy charged particles (solar particles) accelerated during certain solar flares and/or in the

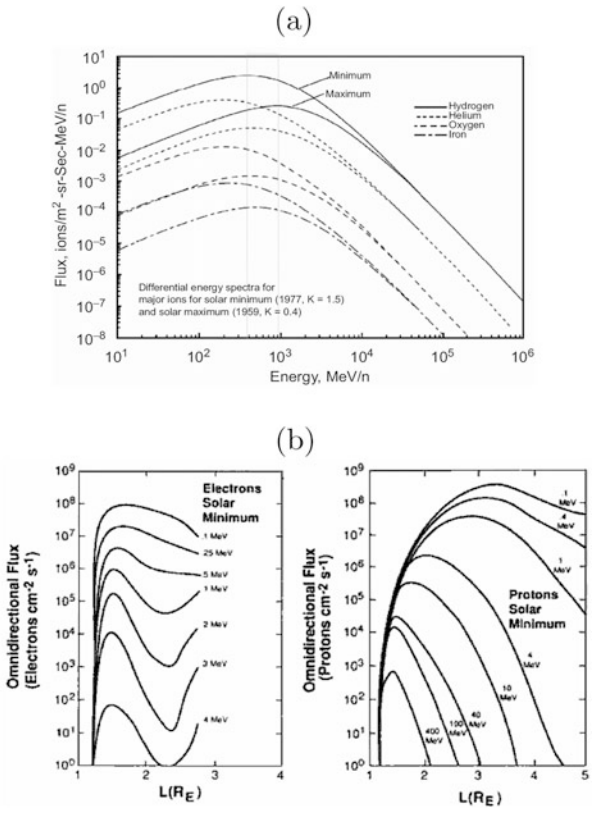


Fig. 21.2 (**a**) Galactic cosmic ray particle spectra and their modification by solar activity [6]. (**b**) Equatorial electron and proton flux vs. the distance from the Earth's center. Each curve gives the total flux above the specified threshold [7]

subsequent coronal mass ejections. These solar particles comprise both protons and heavier ions with variable composition from event to event. Energies typically range up to several hundred MeV and occasional events produce particles of several GeV. Although such events are rare, typically one per month and lasting several hours to days, the flux integrals as large as $10^{10}\,\mathrm{cm}^{-2}$ for protons with energy >1 MeV were measured.

- **Radiation belts;** The charged particles trapped in the Earth's magnetic field form so called Van Allen's belts. The inner belt extends to 2.5 Earth radii and comprises protons up to 600 MeV and electrons up to several MeV. The outer belt extends to 10 Earth radii where there are mainly electrons and soft protons (0.1–100 MeV). The fluxes of electrons and protons trapped in the radiation belts are shown in Fig. 21.2b. The sharp fall of flux at high energies makes shielding very effective.

Environmental Applications

- **Medical application;** The most widely used source of radiation are X-rays, Linacs and radio active isotopes used for cancer treatment. The energy of photons used is: up to 100 keV for X-rays, below 1 MeV for isotopes and up to 25 MeV for Linacs.
- **Fusion in fission reactors and nuclear waste managements;** The main damage comes from neutrons and γ rays, both with energies up to few MeV. Fusion reactors of TOKAMAK type require plasma, fuel impurity and fusion products monitoring instrumentation close to the first wall. The foreseen neutron fluences to which the sensors (e.g. silicon sensors for X-ray spectroscopy) and electronics will be exposed at International Thermonuclear Experimental Reactor (ITER) are comparable with that of the HL-LHC, up to few $10^{16}\,\mathrm{cm}^{-2}$.

21.3 Damage Mechanism in Solid State Detectors and Electronics

As radiation (photons, leptons, hadrons) passes through material, it loses energy by interaction with the electrons and nuclei of the material atoms. The effects produced in the material are dependent on the energy-loss processes and the details of the material structure. The damage in semiconductor detectors can be divided into bulk and surface damage.

21.3.1 Bulk Damage

The interaction with electrons results in creation of electron-hole pairs (ionizing energy loss, ionizing dose) that does not affect the lattice and causes no bulk damage. The bulk damage in crystalline and poly-crystalline material is a consequence of displacement of lattice atoms by impinging particles, due to elastic scattering on a nuclei and nuclear reactions. In order to produce Primary Knocked off Atom (PKA) the transfer of kinetic energy should be sufficient. Approximately 25 eV of recoil energy is required for example in silicon. The displaced atom may come to rest in a interstitial position (I), leaving a vacancy (V) at its original location. If the kinetic energy of the recoiling atom is sufficient ($\sim$5 keV in Si [8]) it can displace further atoms, creating a dense agglomeration of defects at the end of the primary PKA track. Such disordered regions are referred to as defect clusters.

Most of the resulting vacancies and interstitials recombine while others diffuse away and eventually create stable defects with impurity atoms and other vacancies or interstitials. Those defects disturb the lattice periodicity and give rise to energy levels in the band-gap, which alter the properties of the semiconductor. In most semiconductor materials the cross-section for nuclear reaction is much smaller than

Table 21.2 Material properties of some semiconductors used as ionizing particle detectors

Property	Si	Diamond	GaAs	GaN	4H-SiC	a-Si(H)
Z	14	6	31/33	31/7	14/6	14
E_g [eV]	1.12	5.5	1.4	3.39	3.3	1.7
E_{bd} [MV/cm]	0.5	10			2.2–4	
μ_e [cm^2/Vs]	1350	~2000	≤8500	1000	800–1000	1–10
μ_h [cm^2/Vs]	450	~1400	≤400	30	30–115	0.01–0.005
$v_{sat,e}$ [cm/s]	$2 \cdot 10^7$	$2.7 \cdot 10^7$	$1.2 \cdot 10^7$		$2 \cdot 10^7$	
ϵ	11.9	5.5	0.4		9.7	
e–h energy [eV]	3.6	13	4.3	8.9	7.8	4–4.8
e-h/μm for m.i.p.	90	36			51	75
Density [g/cm^3]	2.3	3.5	5.3	6.2	3.2	2.3
Displacement [eV]	25	43	10	Ga-20 N-10		

for elastic scattering, hence the creation rate of defects, resulting from nuclear reactions, is usually more than two orders of magnitude lower when compared to creation rates of defects originating from displaced silicon atoms.

The energy E_p required for an incoming particle of mass m_p to produce PKAs and clusters with a creation threshold E_{th} can be calculated from non-relativistic collision kinematics as

$$E_p = E_{th} \frac{(m_p + m_l)^2}{4 m_p m_l}, \tag{21.2}$$

where the lattice atom has a mass m_l. In silicon a neutron needs at least 175 eV to produce a PKA and 35 keV to form a cluster. For an electron a relativistic kinematics should be used giving 260 keV and 8 MeV. It should be noted that the radiation damage caused by γ-rays from radioactive decays is primarily due to the interaction of Compton electrons with a maximum energy well below the one required for cluster production. The bulk damage is therefore exclusively due to point defects. As the thresholds are of the same order also in other semiconductor materials (see Table 21.2) similar conclusions are valid.

A part of vacancies and interstitials formed immediately after irradiation can recombine, while others diffuse away and eventually recombine or react with other defects or impurities. The defects can evolve in time. They can either dissociate or react with each other and form new defects. The evolution of defects is described by first order dynamics in case of dissociation (Eq. (21.3)) or second order dynamics for reactions of two defects (Eq. (21.4)):

$$X \rightarrow Y \quad , \quad \frac{dN_Y}{dt} = -\frac{dN_X}{dt} = k_1^Y N_Y \tag{21.3}$$

$$X_A + X_B \rightarrow Y \quad , \quad \frac{dN_Y}{dt} = -\frac{dN_{X_A}}{dt} = -\frac{dN_{X_B}}{dt} = k_2^Y N_{X_A} N_{X_B}, \tag{21.4}$$

where $k_{1,2}^Y$ denotes the reaction constants. The Eq. (21.4) turns into a first order process in cases when one type of the reacting defects is present in much larger quantities than the other. The solution of Eq. (21.3) is exponential with

$$N_Y = N_X^0\left(1-\exp\left(-\frac{t}{\tau_1^Y}\right)\right) \quad , \quad N_X = N_X^0 \exp\left(-\frac{t}{\tau_1^Y}\right), \quad \tau_1^Y = \frac{1}{k_1^Y} \tag{21.5}$$

with N_X^0 denoting the initial concentration of defects X proportional to the fluence. The solution of the Eq. (21.4) for $(N_{X_A}^0 > N_{X_B}^0)$ is given by

$$N_Y(t) = N_{X_B}^0 \frac{1 - e^{-k_2^Y t (N_{X_A}^0 - N_{X_B}^0)}}{1 - (N_{X_B}^0/N_{X_A}^0) e^{-k_2^Y t (N_{X_A}^0 - N_{X_B}^0)}} \, . \tag{21.6}$$

In the case of two defects with similar initial concentrations $N_{X_A}^0 = N_{X_B}^0 = N_X^0$ or a reaction between defects of the same type one obtains

$$N_X(t) = \frac{N_X^0}{1 + N_X^0 k_2^Y t} = \frac{N_X^0}{1 + t/\tau_2^Y} \quad , \tau_2^Y = \frac{1}{k_2^Y \, N_X^0} \tag{21.7}$$

$$N_Y(t) = N_X^0 - N_X(t) = N_X^0 (1 - \frac{1}{1 + t/\tau_2^Y}). \tag{21.8}$$

From Eqs. (21.3) and (21.4) it can be seen that for first order reactions, the rate depends linearly on defect concentration while for second order reactions the dependence is quadratic.

Since the energy needed for breaking up the defect (dissociation) or forming a new defect is supplied by the lattice vibrations, the reaction constant is strongly temperature dependent. The lattice atom vibration energy is governed by the Maxwell-Boltzmann distribution. The probability of sufficient energy transfer from lattice vibration to the defect is therefore exponential with temperature (T). If the reaction rate given by the Arrhenius relation is known at T_0 then the rate at T_1 is calculated as:

$$k_{1,2}^Y \propto \exp\left(-\frac{E_a}{k_B \, T}\right) \quad \Longrightarrow \quad \frac{\tau_{1,2}^Y(T_0)}{\tau_{1,2}^Y(T_1)} = \frac{k_{1,2}^Y(T_1)}{k_{1,2}^Y(T_0)} = \exp\left[\frac{E_a}{k_B}\left(\frac{1}{T_0} - \frac{1}{T_1}\right)\right], \tag{21.9}$$

where E_a is the energy required for defect dissociation or formation.

21.3.1.1 Non-Ionizing-Energy-Loss Hypothesis of Damage Effects

The energy loss of impinging particles suffered in a process of displacing lattice atoms is called non-ionizing energy loss—NIEL. First experimental findings have led to the assumption that damage effects produced in the semiconductor bulk by energetic particles may be described as being proportional to non-ionizing energy loss, which is referred to as the NIEL-scaling hypothesis. According to it any displacement damage induced change in the material properties scales with the amount of energy imparted in displacing collisions, irrespective of the spatial distribution of the defects in a PKA cascade and irrespective of the various annealing sequences taking place after the initial damage [10].

The non-ionizing energy deposit in a unit cell of the target nuclei (ρ_{dis}) exposed to the fluence of particles with energy E can be calculated as

$$\rho_{dis} = D(E) \cdot \Phi, \tag{21.10}$$

where $D(E)$ [9] is so-called displacement damage function, sometimes also referred to as damage cross-section. For a spectrum of particles the contributions to the ρ_{dis} for each energy should be summed:

$$\rho_{dis} = \int_0^\infty \frac{d\Phi(E)}{dE} D(E) dE. \tag{21.11}$$

According to NIEL hypothesis ρ_{dis} determines the damage effects. The damage efficiency of any particle spectrum $d\Phi/dE$ can therefore be expressed as that of an equivalent 1 MeV neutron fluence. The equivalent fluence of 1 MeV neutrons Φ_{eq} is calculated as

$$\Phi_{eq} = \kappa \ \Phi = \frac{\rho_{dis}}{D_n(1 \text{ MeV})} \tag{21.12}$$

$$\kappa = \frac{1}{D_n(1 \text{ MeV})} \cdot \frac{\int_0^\infty D(E) \frac{d\Phi}{dE}(E) \, dE}{\int_0^\infty \frac{d\Phi}{dE}(E) \, dE}, \tag{21.13}$$

where κ is so called hardness factor for that particle spectrum and $D_n(1 \text{ MeV})$ the D for 1 MeV neutrons, 95 MeV mb for Si and 10 MeV mb for diamond [12]. The displacement damage cross-section for pions, protons, electrons and neutrons in silicon is shown in Fig. 21.3. The hardness factors for most commonly used irradiation facilities are given in the Table 21.3.

The NIEL hypothesis is violated in silicon for highly energetic charged hadrons. In addition to the hard core nuclear interactions, being dominant for neutrons, charged hadron reactions are also subjected to Coulomb interactions leading to low energy recoils below the threshold for cluster creation. In this case the damage is a mixture of homogeneously distributed point defects and clusters. This distinct

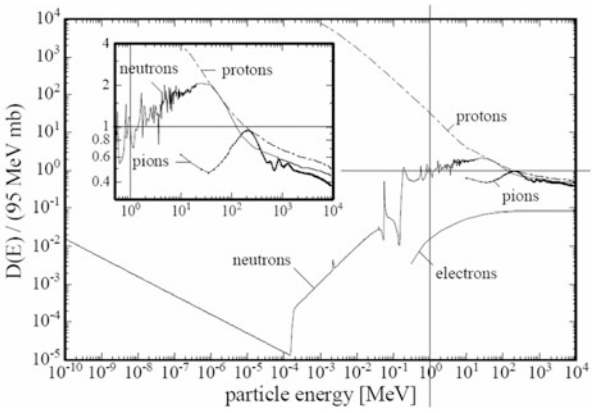


Fig. 21.3 Non Ionizing Energy Loss NIEL for different particles in silicon [11]. The insert shows magnified $D(E)$ for most damaging particles at LHC

Table 21.3 Measured hardness factors of commonly used irradiation particles

	26 MeV[a] protons	70 MeV[b] protons	800 MeV[c] protons	23 GeV[d] protons	200 MeV[e] pions	Reactor [f] neutrons
κ	1.85	1.43	0.71	0.62	1.14	0.92

[a] KIT, Germany and University of Birmingham, UK
[b] CYRIC, Japan
[c] LANL, USA
[d] CERN, Switzerland
[e] PSI, Switzerland
[f] JSI, Slovenia

difference between neutron and proton induced damage is depicted in Fig. 21.4. Different impurities (e.g. O,C) are homogeneously distributed over the volume and the probability for such an impurity to form a defect complex with vacancy or interstitial is much larger if the latter are also homogeneously distributed. Hence, the defects formed after irradiation and consequently the lattice properties can be different for various irradiation particles at equal NIEL. It should be emphasized again that the NIEL scaling can only be regarded as a rough approximation as it disregards the specific effects resulting from the energy distribution of the respective recoils.

21.3.1.2 Impact on Bulk Damage on Detector Performance

As already mentioned the defects in the semiconductor lattice give rise to energy levels (states) in the band gap affecting the operation of semiconductor detector mainly in three ways as shown in Fig. 21.5.

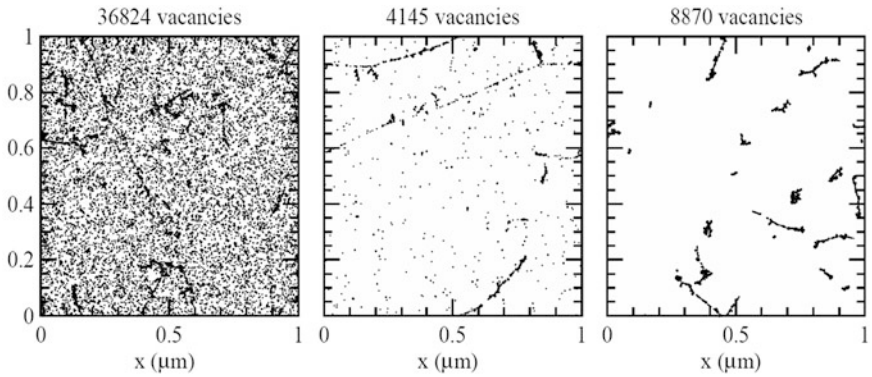


Fig. 21.4 Initial distribution of vacancies produced by 10 MeV protons (left), 23 GeV protons (middle) and 1 MeV neutrons (right). The plots are projections over 1 μm of depth (z) and correspond to a fluence of 10^{14} cm^{-2} [10]

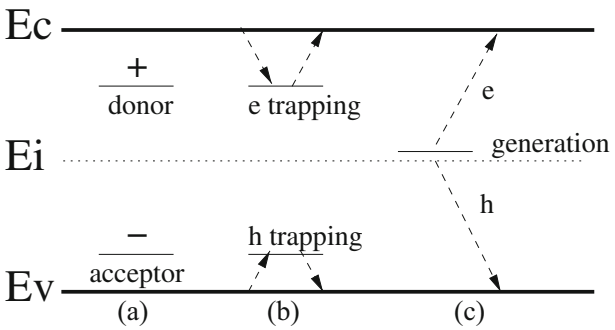


Fig. 21.5 Consequences of deep energy levels to operation of semiconductor detectors: (**a**) charged defects alter the space charge and therefore the electric field, (**b**) defects can trap and detrap free carriers and (**c**) defects act as generation-recombination centers. Electrons and holes are denoted by e and h

- Some of the defects can be charged which leads to (Chap. 5) (Fig. 21.5a) changes in the electric field. For semiconductor detectors this may result in loss of the depleted (active) region requiring an increase of the applied bias. The bias voltage is however limited by the device break down. The space charge is calculated as a difference in concentration of charged donors and charged acceptors,

$$N_{eff} = \sum_{donors} N_t\,(1 - P_t) - \sum_{acceptors} N_t\,P_t, \tag{21.14}$$

where N_t denotes the concentration of deep traps and P_t the probability of a trap being occupied by an electron. The traps continuously emit and capture

carriers. The difference in emission and capture rate is called the excess rate. In a stationary state the occupation probability is constant, therefore excess rates of holes and electrons for a given trap have to be equal. The derivation of occupation probability from this condition can be found in any solid state physics text book. As the P_t is needed for calculation of detector properties we will just state the result:

$$P_t = \left[\frac{c_p\, p + \epsilon_n}{c_n\, n + \epsilon_p} + 1\right]^{-1}, \tag{21.15}$$

$$c_{n,p} = v_{th_{e,h}}\, \sigma_{t_{e,h}}, \epsilon_{n,p} = n_i\, c_{n,p}\, \exp\left(\pm\frac{E_t - E_i}{k_B\, T}\right). \tag{21.16}$$

where $c_{n,p}$ is the capture coefficient and $\epsilon_{n,p}$ emission rate of electrons and holes, respectively. The concentration of free electrons and holes is denoted by n and p and their thermal velocity by $v_{th_{e,h}}$. The capture coefficients and emission rates depend on trap and semiconductor properties. The carrier capture cross-section is given by $\sigma_{t_{e,h}}$ and the level in the band gap by E_t. The Fermi level and free carrier concentration in intrinsic semiconductor are denoted by E_i and n_i. They occupation probability depends on temperature only for levels close to middle of the band-gap. The exponential term in Eq. (21.15) prevails once E_t is few $k_B\, T$ away from the E_i. If follows from here that only donors in the upper part of the band gap and acceptors in lower part of the band gap contribute to the space charge.

- The states can act as trapping centers for the drifting charge generated by the particles we want to detect (Fig. 21.5b). If trapped charges remain trapped and do not complete the drift within the integration time of the read-out electronics they are lost for the measurement, which leads to smaller signal.

  The probability for electrons and holes to be trapped at the trap t can be calculated as

$$\frac{1}{\tau^t_{tr_e}} = c_n\,(1 - P_t)\, N_t, \quad \frac{1}{\tau^t_{tr_h}} = c_p\, P_t\, N_t. \tag{21.17}$$

  The trapping time $\tau^t_{tr_{e,h}}$ represents the mean time that a free carrier spends in the part of the detector before being trapped by t. According to Eq. (21.17) electron traps have energy levels in the upper part of the band gap ($P_t \approx 0$), while hole traps have energy levels in the lower part of the band gap ($P_t \approx 1$).

  To get the effective trapping probability $1/\tau_{eff_{e,h}}$ for electrons and holes one has to sum over the trapping probabilities of all traps with emission times $(1/\epsilon_{n,p})$ longer than integration time of the electronics:

$$\frac{1}{\tau_{eff_e}} = \sum_t^{defects} c_n\,(1 - P_t)\, N_t, \tag{21.18}$$

$$\frac{1}{\tau_{eff_h}} = \sum_{t}^{defects} c_p \, P_t \, N_t. \tag{21.19}$$

The emission times decrease with distance from the mid-gap and become at certain energy level short enough not to be included in the Eq. (21.19). The traps close to the mid-gap have therefore a dominant contribution to the effective trapping times.

- States close to the mid-gap region also act as generation-recombination centers (Fig. 21.5c). The thermally generated electron hole pairs are separated in the electric field before they can recombine, which gives rise to the bulk generation current. The increase of current leads to the increase of noise and power dissipation.

  The generation current can be calculated with the assumption of equal generation rates $G_t = G_n = G_p$ of electrons and holes in thermal equilibrium:

$$G_t = N_t \, P_t \, \epsilon_n = N_t \, \frac{\epsilon_n \, (\epsilon_p + c_n \, n)}{\epsilon_n + \epsilon_p + c_p \, p + c_n \, n} \tag{21.20}$$

$$G_t = N_t \, \frac{1}{1/\epsilon_n + 1/\epsilon_p} \quad \text{for} \quad n, \, p \approx 0. \tag{21.21}$$

  Both carrier types generated in the active volume drift to the opposite electrodes. The current density, albeit different for holes and electrons, is constant everywhere in the detector. The measured current is therefore calculated as

$$I = e_0 \, w \, S \sum_{t}^{defects} G_t \tag{21.22}$$

  where w denotes the active thickness and S the active surface of the detector. It follows from Eq. (21.21) that only the levels close to mid-gap $E_i \sim E_t$ contribute significantly to the current. If traps are far from the mid-gap, emission times are either very long or very short.

Apart from the changes in the depletion region, the properties of the non-depleted silicon bulk are also affected by irradiation. The resistivity of the bulk increases. The increase depends on both initial dopant concentration as well as on irradiation fluence. The minority carrier lifetime also decreases as $1/\tau_r \propto \Phi$ and reaches values of few tens ns at $\Phi = 10^{14}\,\mathrm{cm}^{-2}$ and below ns at $\Phi > 10^{16}\,\mathrm{cm}^{-2}$ [13].

Recent measurements [14] also show that mobility of free carriers is affected by radiation. The concentration of defects, not only electrically active, is high enough to affect the low field mobility. A significant decrease of low field mobility was observed at fluences of $\Phi_{eq} > 5 \cdot 10^{15}\,\mathrm{cm}^{-2}$.

Although silicon detectors are by far the most widely used there are other semiconductor detectors which can be used in high radiation fields and have a higher

PKA displacement energy. The material properties of different semiconductors used as particle detectors are summarized in Table 21.2.

Effects of irradiation on detector performance strongly depend on the choice of material. In wide band gap semiconductors for example the rate of thermally generated carriers will be small even if states close to mid-gap are present in abundant concentrations due to small intrinsic carrier concentration. Thus the leakage current increase is negligible. If the drift velocity is large and charge collection time is short then the increase of trapping probability will be less important. The small dielectric constant reduces the capacitance of a detector leading to lower noise, which can partially compensate for larger e-h pair creation energy. The choice of the semiconductor detector for a specific application is often governed by a compromise in semiconductor properties. Also availability, reliability and experience play an important role. In this respect diamond is the choice of detector material next to silicon.

21.3.1.3 Most Important Defects in Silicon

A lot of effort was invested over the R&D phases of LHC/HL-LHC in identifying the defects responsible for changes in performance of silicon detectors. A comprehensive list of defects identified by so called "microscopic" techniques such as Deep Level Transient Spectroscopy (DLTS) or Thermally Stimulated Current (TSC) can be found in [15]. The summary plot with the most important defects is shown in Fig. 21.6. The effects for which they are mainly responsible will be addressed in

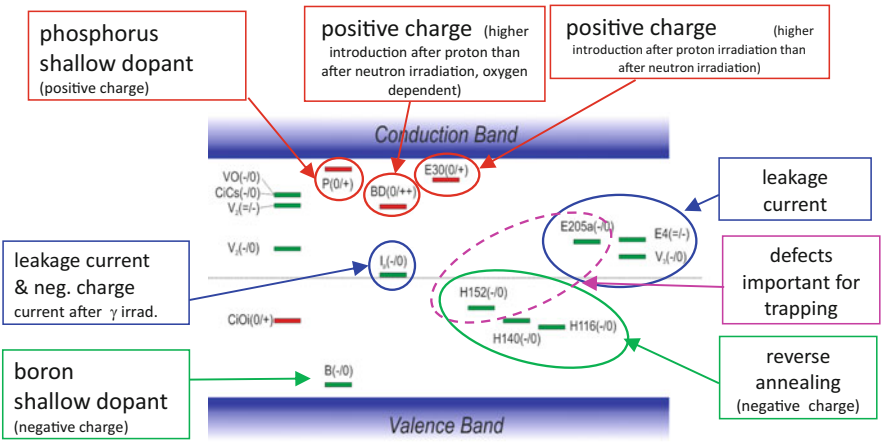


Fig. 21.6 A schematic view of known defects and their main effects on the detector performance. The defect charge state is given in brackets. For the defects with unknown chemical composition the temperature at which electron -E or hole-H traps were identified with DLTS/TSC techniques is used. The near mid-gap H levels are likely multivacancy complexes

the following sections. Note, that for only few identified energy levels the chemical composition of the corresponding defects is known.

21.3.2 Surface Damage

The semiconductor detector bulk needs to form a contact with readout electronics. The contacts used, either Ohmic or Schottky, as well as the rest of the surface are prone to changes due to irradiation. The description of surface radiation damage given here will be focused on the border of silicon bulk and oxide (Chap. 5). The surface damage affects the electrical properties of the detectors such as inter-electrode resistance, inter-electrode capacitance and dark current. It is particularly important for sensors where charge flow is close to the surface, such as 3D-Si detectors, CCDs, Active CMOS Pixel Detectors and MOS-FET transistors.

The surface of particle detectors is usually passivated by thermal oxidation [16]. The oxide isolates and stabilizes the crystal surface with respect to chemical and electrical reactivity. The cross-section of the device surface is generally divided into silicon/oxide interface and oxide bulk depicted in Fig. 21.7. The border region between oxide and silicon crystal is characterized by a large defect density due to bond stress. In general surface defects can be caused by growth and irradiation. According to their position in the oxide the traps are divided in the oxide bulk traps (OT), border traps (BT) and interface traps (IT). The latter two are located close to the interface and can exchange charges with underlying silicon (switching traps). The oxide traps are mostly donors, which is the reason that net oxide charge

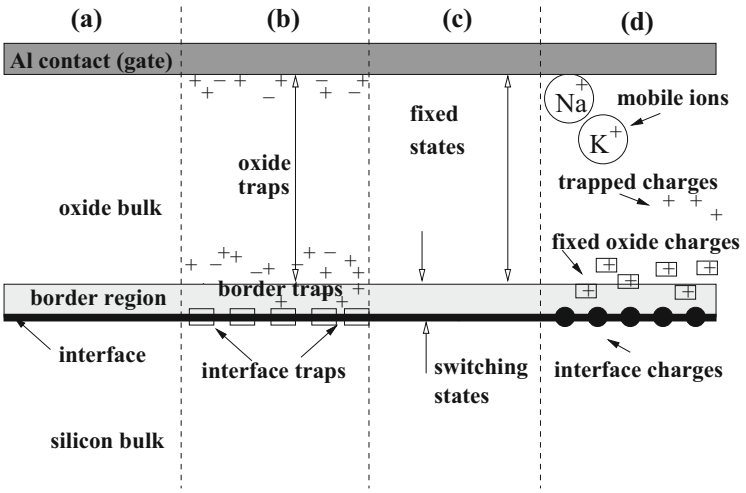


Fig. 21.7 Schematic view of the surface of a silicon detector according to [17]; (**a**) surface regions (**b**) trap locations (**c**) states (**d**) oxide charges

density is always positive. The most important oxide defects are trivalent Si (≡ Si·, donor), interstitial oxygen (O_I, donor) and non-bridging oxygen (≡ Si–O·, acceptor). Other important defects include hydrogen related defects (all donors) [18]. Hydrogen is particularly important since it passivates the dangling bonds by attaching to them. The build-up of interface traps is not fully understand yet and there are different models explaining it [18, 19]. The bulk and interface traps formed during processing of the oxide can be passivated by annealing (350–500 °C) in hydrogen rich environment.

If the creation of e–h pair in the silicon bulk is completely reversible process, it is not in SiO_2 and at the interface. Ionizing radiation has a significant impact on the defect generation and activation. The damage mainly manifests itself as a regeneration process of already present but deactivated defects. Hence the processing of the oxide, preparation and temperature treatments (annealing) impacts the performance after irradiation.

Although the underlying physics of formation is not yet fully understood, it is assumed that radiation ionizes oxide bulk defects that remain charged

$$\mathrm{Si}\cdot + \mathrm{radiation} \rightarrow \mathrm{Si}^{+} + \mathrm{e}^{-} \tag{21.23}$$

or free holes are trapped by passivated defects

$$\mathrm{O_I} + \mathrm{h} \rightarrow \mathrm{O_I}^{+}. \tag{21.24}$$

Similarly to oxide bulk damage the interface state density also increases with irradiation. After [20] the interface states are generated by breaking up the bonds between surface silicon atoms (Si_s) and hydrogen, due to hole trapping at the interface (Si_s-H+h→Si_s·+H^+; Si_s-H+h→Si_s^++H· followed by Si_s^++e^- →Si_s·). The dangling bonds enable surface silicon atoms to react with the underlying silicon and induce different states in the silicon band-gap. The state build-up can continue over a long period of time after exposure to radiation.

The electrons are much more mobile in the oxide ($\mu_e(20\,°\mathrm{C}) \sim 20\,\mathrm{cm}^2/\mathrm{Vs}$) and are in the presence of electric field promptly swept away, while holes ($\mu_h(20\,°\mathrm{C}) = 10^{-4}-10^{-11}\,\mathrm{cm}^2/\mathrm{Vs}$) slowly drift to the interface. The absence of electric field in the oxide is therefore beneficial as the recombination can take place in the oxide bulk as well as at the interface.

21.3.2.1 Impact of Surface Damage on Device Properties

Positive Oxide Charge

As shown by many experiments the exposure to ionizing radiation causes an increase of positive space charge. The different contributions to the oxide charge are shown in Fig. 21.7. Apart from the oxide traps and mobile ion impurities also trapped holes at interface states contribute to the positive oxide charge. An

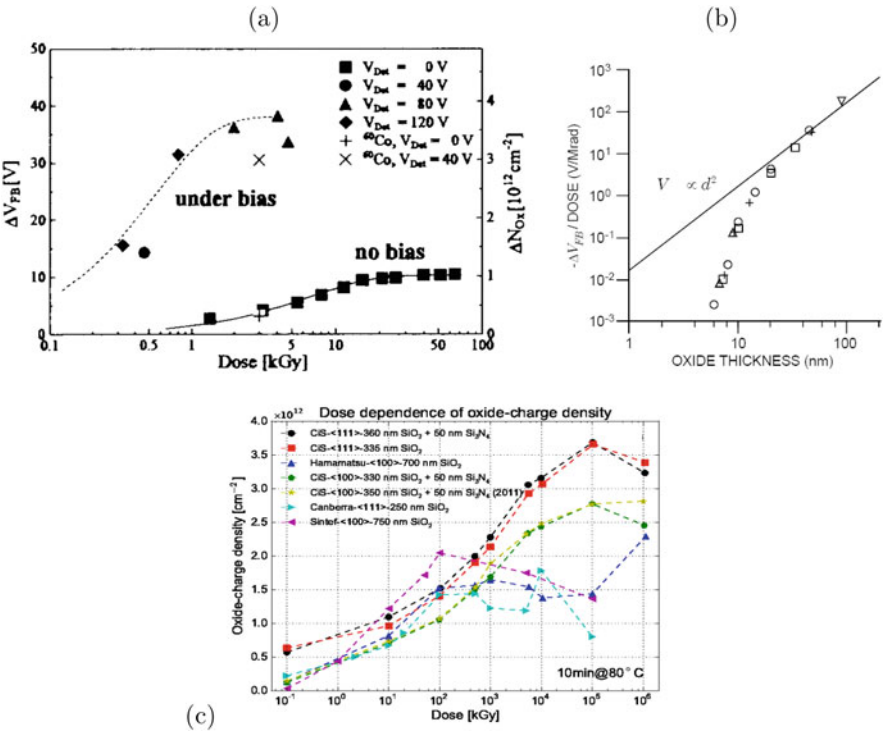


Fig. 21.8 (**a**) Oxide charge measured from a change in flat band voltage for silicon gated diodes [21] after irradiation with 20 keV electron and γ-rays from ^{60}Co. (**b**) Dependence of flat band voltage on oxide thickness [22]. (**c**) Recent measurements to very large doses for samples with different producer/orientation/oxide thickness [23].

effective net sheet charge (surface density) in the oxide N_{ox} is calculated as the sum of all contributions. It has been shown that under bias the oxide charge density increases with irradiation up to few kGy where it starts to exhibit saturation. In an unbiased devices saturation occurs at significantly larger doses up to few 10 kGy (see Fig. 21.8) [21]. The saturation sheet charge depends on thickness of the oxide and is of order $N_{ox} = 10^{12}\,\mathrm{cm}^{-2}$. Latest measurements show an increase of oxide charge, although at a much slower rate, up to the doses of 1 GGy (see Fig. 21.8c).

The positive oxide charge attracts electrons which can form a conductive layer underneath the surface. The resistivity between the nearby n^+ contacts can therefore decrease producing a short circuit. A p^+ implant is therefore commonly used to cut these conductive paths. A more novel approach is to use a moderate p implant over the whole surface (p-spray [24]). The p-spray dose must be sufficiently high ($\approx 10^{11}-10^{13}$ ions/cm^2; the same order as N_{ox}) to prevent decrease of inter-strip resistivity and not too high to cause early breakdowns. Very often both methods are used together.

In very thin oxides the tunneling of electrons from nearby electrodes occurs. The oxide traps get passivated, by reversing the reactions described by Eqs. (21.23), (21.24). Thinning down the oxide therefore reduces the N_{ox} (see Fig. 21.8b) [22], which makes the device more radiation hard. The flat band voltage which should follow the $V_{FB} \propto d^2$, if the oxide charge is uniform, shows a steep decrease in thin oxide films $<$20 nm. The importance of this effect will be discussed in section on radiation hard electronics.

Surface Generation Current

Interface states act as charge carrier generation centers. As soon as the silicon surface is depleted, the thermally generated carries are separated in electric field and contribute to the dark current of a nearby $p-n$ junction or a MOS transistor. This current is called interface generation current and is calculated as

$$I_{ox} = e_0 \, n_i \, S_s \, v_{surf} \tag{21.25}$$

where v_{surf} is the surface recombination velocity and S_s the depleted silicon surface area. The surface recombination velocity is directly proportional to the density of interface states. The density of states rather than discrete states is used as experimentally it is impossible to distinguish between different trap levels [25]. The increase of surface current and surface recombination velocity with irradiation is shown in Fig. 21.9.

Trapping

The interface states act as trapping centers for the charge drifting close to silicon surface in analogous way to trapping of drifting carries in the bulk. Equation 21.17 is multiplied with an exponential term $\exp\left(\frac{e_0 < \psi >}{k_B T}\right)$ to take into account the average band bending $< \psi >$ close to the surface.

21.4 Detector Technologies

21.4.1 Design Considerations

The design of the detector should minimize the radiation effects most crucial for the successful operation of the detector while retaining the required functionality. The material and operational conditions determine to a large extent the radiation hardness of a detector. However, some of the radiation effects can be reduced by

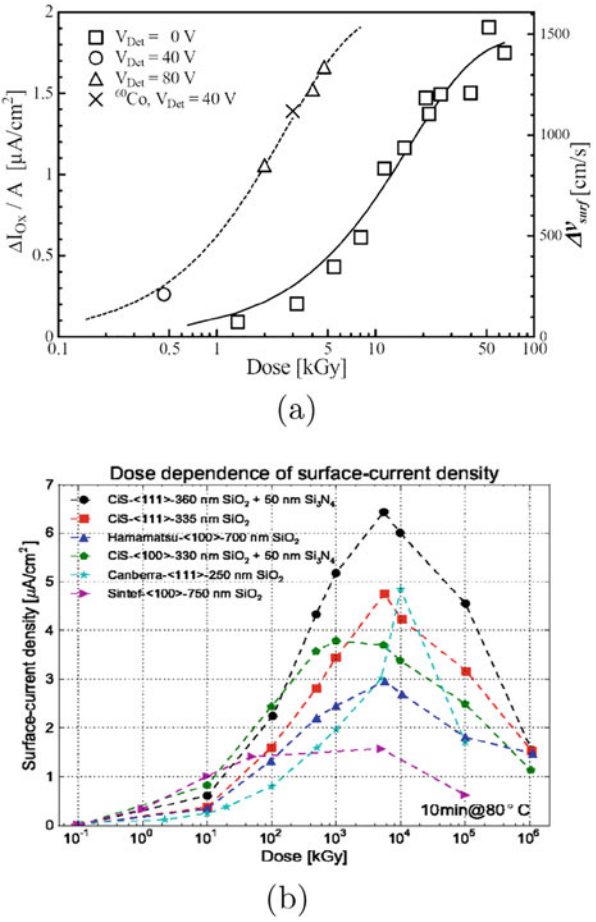


Fig. 21.9 (**a**) The increase of surface current density (surface recombination velocity) after 20 keV electron and ^{60}Co irradiations for biased and unbiased gate. (**b**) Surface current density after 12 keV X-rays irradiations of different samples to very high doses [23]

a choice of the read-out electrodes and detector geometry. At the new accelerator experiments the largest obstacle is the radiation-provoked decrease of measured charge and increase of noise. The consequent degradation of signal-to-noise ratio can lead to the loss of detection efficiency up to the level where successful operation of the detectors is no longer possible.

In terms of charge collection the radiation hard detector design follows directly from the calculation of the induced charge Q. The current induced (I) by a motion of charge q in the detector is given by Shockley-Ramo's theorem [26] and is discussed in the section on signal processing. The charge induced in the electrodes is given

by the difference in the weighting potential (U_w) traversed by the drifting charge (Chap. 10, Eq. 10.2):

$$Q(t) = q[U_w(\vec{r}(t)) - U_w(\vec{r_0})], \tag{21.26}$$

where $\vec{r}_0$ and $\vec{r}$ denote position at the both ends of the traversed path. The distinct difference in weighting potential for a pixel detector and simple pad detector is shown in Fig. 21.10 and discussed in section 6.2.2.

For an electron hole pair the induced charge is a sum of both contributions $Q_{e-h} = Q_e + Q_h$. A track of an ionizing particle therefore induces the charge Q^t

$$Q^t = \sum_{all\ pairs} Q_e + Q_h = Q_e^t + Q_h^t, \tag{21.27}$$

$$Q_{e,h}^t = \mp e_0 \sum_i^{e,h} U_w(\vec{r}_i) - U_w(\vec{r}_{i,0}). \tag{21.28}$$

If all carries complete the drift on the sensing electrode $U_w(\vec{r}_i) = 1$ if on non-sensing $U_w(\vec{r}_i) = 0$. In the absence of trapping and homogeneous ionization the sum in Eq. (21.28) becomes integral which can be easily calculated. For the track through the center of the pixel shown in Fig. 21.10 the contribution of electrons drifting to sensing electrode is $Q_e^t/Q^t = 0.82$, which is significantly larger than $Q_e^t/Q^t = 0.5$ for pad detectors. The fact that in segmented devices one carrier type contributes more to the total induced charge, can have important consequences after irradiation if the difference in mobility or/and trapping probability is large for electrons and holes.

If carriers are trapped and not released in time to finish the drift within the integration time of the amplifier (t_{int}) then $U_w(r_i) \neq 1, 0$. Using $\vec{v_{e,h}} = \mu_{e,h}\vec{E}$ and $q = e_0 \exp\left(\frac{-t}{\tau_{eff,e,h}}\right)$ the Eqs. (21.28), Eq. 1 (Section 6) turn to

$$Q_{e,h}^t = \mp e_0 \sum_i^{e,h} \int_0^{t_{int}} \exp\left(\frac{-t}{\tau_{eff,e,h}}\right) \mu_{e,h}(E) \left[\vec{E}(\vec{r}_i) \cdot \vec{E_w}(\vec{r}_i)\right] dt, \tag{21.29}$$

where $\mu_{e,h}$ represents carrier mobility. Three conclusions can be drawn without actually solving the Eq. (21.29) for a given detector and charge particle track:

- A better charge collection efficiency CCE (ratio of measured and generated charge) of the hit electrode is achieved when it collects the carriers with larger $\mu \cdot \tau_{eff}$. They contribute a larger part to Q^t and hence reduce the effect of the trapping.
- If the electric field can not be established in the entire detector (e.g. partial depletion or polarization of detector) it is important to have the region with

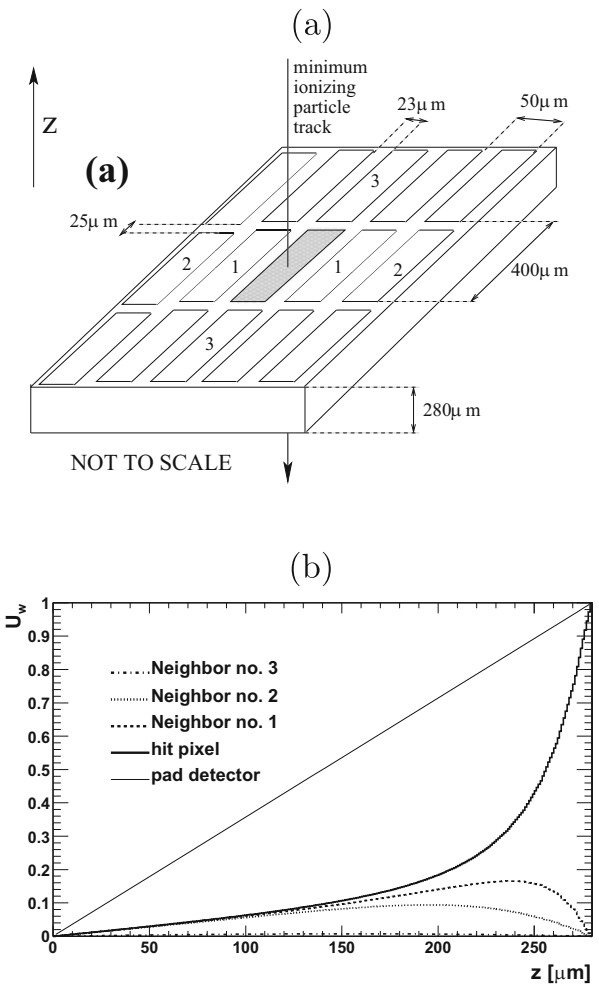


Fig. 21.10 (**a**) A schematic picture of the ATLAS pixel detector with pixel dimensions of $400 \times 50\,\mu\text{m}^2$. The hit pixel for which the U_w was calculated is shaded. Neighbors are denoted by the corresponding numbers. (**b**) The weighting potential along the axis through the center of the hit pixel and through the center of the three closest neighbors. For comparison U_w of a pad detector is also shown

electric field around the read-out electrodes, where E_w is large (large $\vec{E} \cdot \vec{E_w}$). Operation of partially depleted detectors therefore requires that the junction grows from the segmented side. Growth of depletion region from the back of the detector, shown in Fig. 21.10, would result in smaller induced charge in hit pixel than expected from the thickness of the active region.

- A detector design where the number of generated e–h pairs is disentangled from their drift time is optimized for large induced charge.

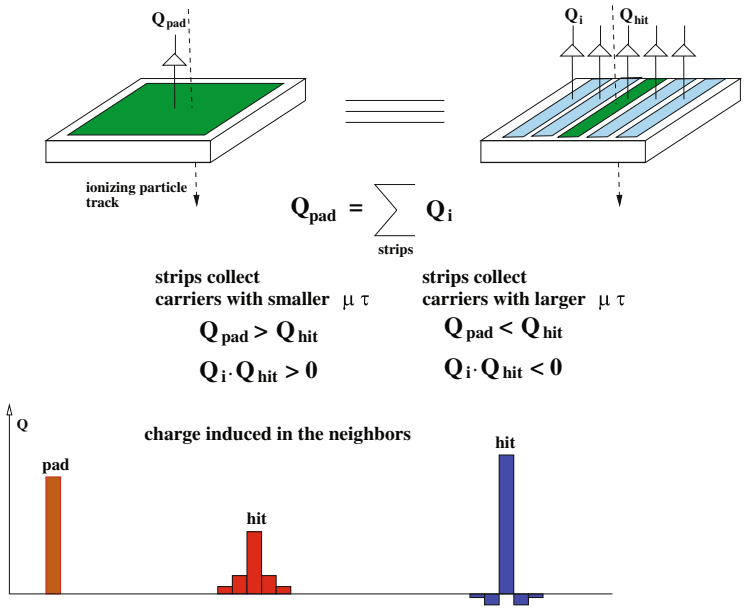


Fig. 21.11 Explanation of trapping induced charge sharing

As U_w depends on the geometry only it is obvious that it is possible to optimize the electrode design for maximum signal. However, as the paramount parameter for any detector is its signal-over-noise ratio, the optimization should also include the inter-strip capacitance and leakage current of electrode both affecting the noise.

Charge collection in segmented devices leads to "signal cross-talk" as described in the section on Signal processing 6.2.1. The bi-polar current pulses in the neighboring electrodes (see e.g. U_w in Fig. 21.10) yield zero net charge for integration time larger than the drift time (see Signal processing Fig. 6.2). In irradiated detector some of the carriers are trapped and do not complete their drift. Therefore the integrals of the bipolar pulses do not vanish. A significant amount of charge can appear in the neighbors adding to the usual charge shared by diffusion (see explanation in Fig. 21.11). Unlike diffusion, where the polarity of the induced charge is equal for all electrodes, the trapping can result in charges of both polarities. If electrodes collect carriers with smaller $\mu \cdot \tau_{eff}$, the polarity of the charge is the same for all electrodes. Otherwise the polarity of the charge induced in the neighbors is of opposite sign compared to the hit electrode [27, 28]

The effect can be used to enhance the spatial resolution due to larger charge sharing at the expense of smaller charge collection efficiency or vice versa.

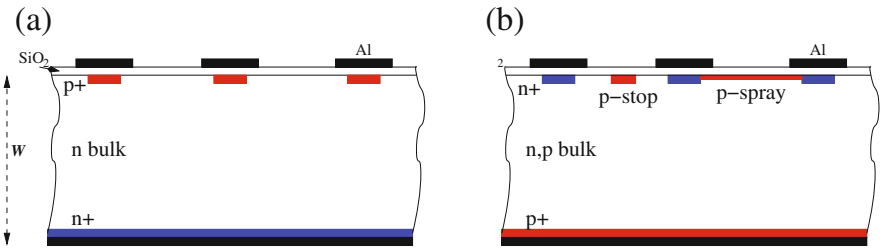


Fig. 21.12 Schematic view of (**a**) $p^+ - n - n^+$ and (**b**) $n^+ - n - p^+$, $n^+ - p - p^+$ strip detectors (AC coupled)

21.4.2 Silicon Detectors

Silicon is by far the most widely used semiconductor detector material. A large majority of silicon particle detectors exploit the asymmetric $p - n$ junction bias in the reverse mode as a basic element. Up to recently the detector grade silicon was produced by the so called float zone (FZ) technique, where concentration of impurities and dopants can be precisely controlled to very low values ($\sim 10^{11}$ cm^{-3}). The step further in radiation hardening of silicon detectors was the enrichment of the float zone silicon through oxygen diffusion (DOFZ). Recently, detectors were processed on Czochralski[1] and epitaxially grown silicon and are in some respects radiation harder than float zone detectors.

Most of the detectors used up to now were made on n-type silicon with p^+ readout electrodes (see Fig. 21.12a), which collect holes. Electrons have larger $\mu\,\tau_{eff}$ in silicon, hence n^+ readout electrodes are more appropriate for high radiation environments where the loss of charge collection efficiency is the major problem. They are mostly realized by segmentation of n^+ side of the n-type bulk (see Fig. 21.12b), which however requires more complex processing on both detector sides. The double sided processing can be avoided by using p-type bulk material with n^+ electrodes [29]. This is the preferred choice silicon detector type at HL-LHC.

21.4.2.1 Effective Doping Concentration

The defects produced by irradiation lead to change of the effective doping concentration. The main radiation induced defects responsible for the change of effective dopant concentration can be found in Fig. 21.6 and consist of both donors and acceptors.

[1]If magnetic field is used to control the melt flow in crucible the process is called Magnetic-Czochralski.

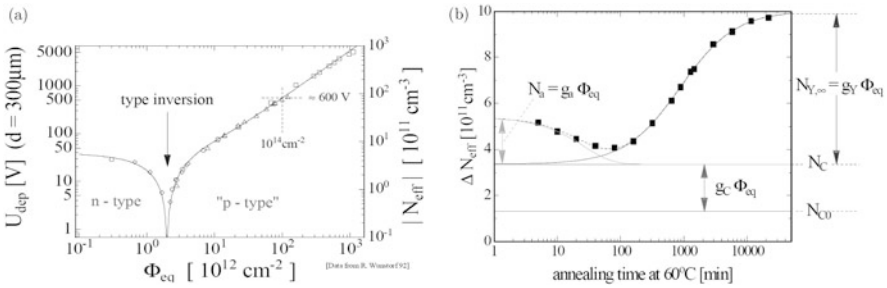


Fig. 21.13 (**a**) Effective doping concentration in standard silicon, measured immediately after neutron irradiation [30] (**b**) Evolution of ΔN_{eff} evolution with time after irradiation [31]

It is a well established, that irradiation by any particle introduces effectively negative space charge in detectors processed on float zone silicon, which is most commonly used. The change in effective doping concentration is reflected in the full depletion voltage V_{fd}, needed to establish the electric field in the entire detector:

$$V_{fd} = \frac{e_0 \, |N_{eff}| \, W^2}{2\epsilon_0 \epsilon}. \tag{21.30}$$

The V_{fd} of initially n-type detectors ($p^+ - n - n^+$), therefore decreases to the point, where the negative space charge prevails, so called space charge sign inversion point (SCSI). The N_{eff} turns to negative and depleted region grows from the n^+ contact at the back. The V_{fd} thereafter continues to increase with fluence beyond any tolerable value, which is usually set by the breakdown of a device (see Fig. 21.13a). The space charge of p-type detectors ($n^+ - p - p^+$) remains negative with irradiation so that the main junction stays always at the front $n^+ - p$ contact.

For both detector types not only deep radiation induced defects are created, but also initial shallow dopants are electrically deactivated (removed)—so called initial dopant removal. The initial dopant removal impacts to large extent the performance of some detector technologies such as Low Gain Avalanche Detectors and depleted CMOS detectors, which will be reviewed later.

Evolution of Effective Dopant Concentration—Hamburg Model

After the irradiation the defects responsible for space charge evolve with time according to defect dynamics described by Eqs. (21.3), (21.4). The time scale of these processes varies from days to years already at close to room temperatures which makes the annealing studies lengthy procedures. At elevated temperature the underlying defect kinetics can be accelerated, and thus the simulation of the damage investigation at real experiments spanning several years is possible in weeks.

The radiation induced change in the effective doping concentration is due to historical reasons defined as $\Delta N_{eff} = N_{eff,0} - N_{eff}(t)$, where $N_{eff,0}$ denotes the initial doping concentration. The fact that the radiation introduced space charge is negative means that ΔN_{eff} is positive. The evolution of N_{eff} after irradiation is shown in Fig. 21.13b. ΔN_{eff} initially decreases, reaches its minimum and then starts to increase. The measured evolution can be described by a so called Hamburg model, which assumes three defects [31] all of them obeying first order kinetics (see Eq. (21.3)). The initial decrease of ΔN_{eff} is associated with decay of effective acceptors (N_a). After a few days at room temperature a plateau, determined by defects stable in time (N_c), is reached. At late stages of annealing effective acceptors are formed again (N_Y) over approximately a year at room temperature. The corresponding equations are:

$$\Delta N_{eff} = N_{eff,0} - N_{eff} = N_a(\Phi, t) + N_c + N_y(\Phi, t) \tag{21.31}$$

$$\Delta N_{eff} = g_a \Phi_{eq} \exp\left(-\frac{t}{\tau_a}\right) + N_c + g_Y \Phi_{eq}(1 - \exp\left(-\frac{t}{\tau_{ra}}\right)) \tag{21.32}$$

$$N_c = \pm N_{id}(1 - \eta(1 - \exp(-c \cdot \Phi_{eq}))) + g_c \Phi_{eq}, \tag{21.33}$$

where g_a, g_c and g_Y describe the introduction rates of defects responsible for the corresponding part of the damage and τ_a and τ_{ra} the time constants of initial and late stages of annealing.

The **stable part** of the damage incorporates also **initial dopant removal**, where $\pm N_{id}$ (negative/positive sign for donors/acceptors) denotes the concentration of initial dopants, η fraction of removed dopants and c the removal constant. Displacement of the initial dopant from the lattice site, deactivates it. Once in the interstitial position, initial dopants (mainly boron and phosphorous) can react with other defects leading to possibly new electrically active defects. The new defects formed can also be charged, hence the removal can be partial, i.e. $N_{id} \neq N_{eff,0}$ [32, 33]. For example, the interstitial boron can undergo different reactions with impurities forming both donor and acceptor like defects [32]. As the reactions can take place also with impurities the removal rate depends on their concentration.

The initial donor (phosphorous) removal was intensively studied for high resistivity $p^+ - n - n^+$ detectors [34], where initial donor removal is attributed to formation of electrically inactive Vacancy-Phosphorous (V-P) complex. The rate of removal was found to depend on initial concentration with $N_{id} \times c \approx 0.008\,\mathrm{cm}^{-1}$. The reason for such relation is unclear. It was observed that donor removal is complete for charge hadron irradiated detectors while around half of the initial donors remain effectively active after neutron irradiations ($\eta \sim 0.45 - 0.7$).

The initial acceptor (boron) removal was much less studied in the $n^+ - p - p^+$ particle detectors, more for solar cells [35]. The required radiation hardness of p-type detectors for HL-LHC is such that deep acceptors exceed the concentration of

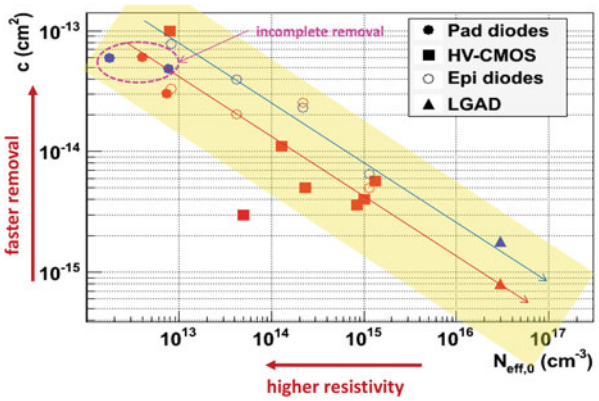


Fig. 21.14 Initial acceptor removal rate dependence on initial dopant concentration. The data were obtained from measurements with different detectors/technology: pad diodes (float zone and epitaxial), depleted (HV) CMOS and LGADs. The red markers show neutron irradiations and the blue markers show fast charged hadron irradiations. The red and blue arrows guide the eye. Data from Refs. [36–41]

Table 21.4 The survey of Hamburg model parameters for standard and diffusion oxygenated float zone detectors

	Standard FZ		Diffusion Oxygenated FZ	
	Neutrons	Charged hadrons	Neutrons	Charged hadrons
g_a [cm^{-1}]	0.018	–	0.014	–
τ_a [h at 20 °C]	55	–	70	–
g_c [cm^{-1}]	0.015	0.019	0.02	0.0053
g_Y [cm^{-1}]	0.052	0.066	0.048	0.028
τ_{ra} [days at 20 °C]	480	500	800	950

The uncertainty in the parameters is of order 10% and mainly comes from variation of silicon materials used

initial ones by far, hence their removal was not in focus. However, new detector technologies (LGAD, depleted CMOS) with significant/dominant concentration of initial dopants also after foreseen fluences, triggered extensive studies of initial acceptor removal. Similarly to donor removal c was found to depend on initial concentration as shown in Fig. 21.14. The rate of removal is around two times larger for fast charged hadrons and only for large initial dopant concentrations the removal is complete ($\eta \approx 1$).

The **parameters of the Hamburg model** related to radiation induced defects (deep traps) are given in the Table 21.4 and are valid for p- and n-type silicon detectors. For reasons that will be explained later, the model parameters are also shown for FZ detectors which were deliberately enriched by oxygen.

The time constants of initial (τ_a) and late stage annealing (τ_{ra}) can be scaled to different annealing temperatures by using Eq. (21.9). The activation energies for initial and long term annealing are $E_{ra} \approx 1.31\,\text{eV}$ and $E_a \approx 1.1\,\text{eV}$ [34].

After around 80 min annealing at 60 °C $N_a, N_y \ll N_c$ and ΔN_{eff} is almost entirely due to stable defects. If the initial dopant removal is complete or initial dopant concentration is small (with respect to to deep defects) the effective doping concentration is given by a simple relation $|N_{eff}| \approx g_c \cdot \Phi_{eq}$.

Often the irradiations follow the planned operation scenario. For example at LHC the detectors are operated at $T \approx -10°\text{C}$ for 1/3 of the year then stored for few weeks at close to room temperature and the rest of the year at $T \approx -10°\text{C}$. The corresponding temperature history of a whole year can be compressed roughly to 4 min at 80°C. The whole period of operation therefore consists of multiple irradiation and annealing steps, which is also referred to as CERN scenario [34].

The parameters of Hamburg model are used to predict the evolution of full depletion voltage of silicon pixel ($n^+ - n - p^+$) and strip detectors ($p^+ - n - n^+$) at LHC experiments. The agreement of predictions with measurements during LHC operation was good, as shown on few examples in Fig. 21.15.

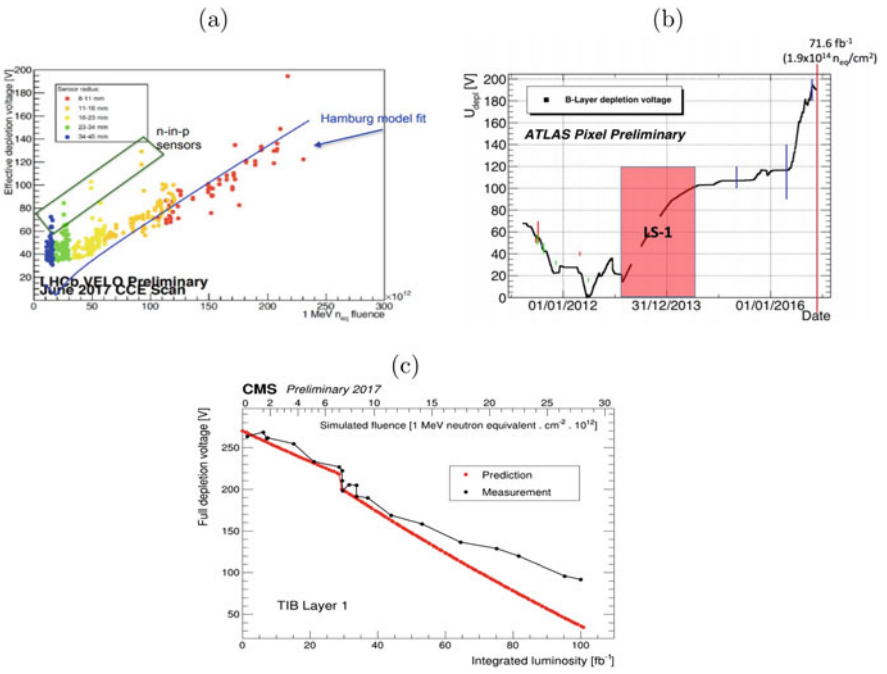


Fig. 21.15 The agreement of predicted and measured V_{fd} for (**a**) LHCb Velo detector [42], (**b**) ATLAS-Insertable B layer pixel detector [43] (**c**) CMS—strip detectors in the outer region [44]. For (**b**) the prediction is denoted by black dots and measurements as bars with different colors

As can be seen in Figs. 21.15, the agreement of Hamburg model with measurements is reasonable and allows for predictions of operation up to the end of their lifetime at LHC. It is evident that careful planning of maintenance and technical stops is required to keep V_{fd} as low as possible. Even though oxygen rich silicon was used for ATLAS pixel detectors, they will be operated under-depleted at least for some time at the end of LHC operation. The depleted region after space charge inversion grows from the pixel side and for $V_{bias} < V_{fd}$ the detector performance is similar to that of somewhat thinner detector, still providing efficient tracking.

On the other hand irradiated strip detectors at LHC ($p^+ - n - n^+$) require at all times $V_{bias} > V_{fd}$ as the region around the strips needs to be depleted for achieving sufficient charge collection efficiency. The maximum bias voltage for e.g. ATLAS strip detectors is set to 450 V, which is sufficient for full depletion over the entire operation program before the HL-LHC upgrade. Standard float zone detectors are used for the fact that the larger fraction of damage is coming from neutrons and oxygenated detectors would therefore offer no significant advantage.

Defect Engineering

The radiation tolerance of silicon can be improved by adequate defect engineering. Defect engineering involves the deliberate addition of impurities in order to reduce the radiation induced formation of electrically active defects or to manipulate the defect kinetics in such a way that less harmful defects are finally created. It has been established that enhanced concentration of oxygen in FZ detectors reduces the introduction rate of stable defects by factor of $\sim$3 after charged hadron irradiations (see Table 21.4). The most likely explanation is that oxygen acts like a trap for vacancies (formation of an uncharged V-O complex) and therefore prevents formation of charged multi-vacancy complexes. In addition, Oxygen is also related to formation of deep donors (see Fig. 21.6).

On the opposite carbon enhances the concentration of vacancies as it traps interstitial silicon atoms and reduces the recombination. Since the concentration of oxygen is not high enough in the disordered regions-clusters, it has little or no effect after neutron irradiations. Different stable damage in neutron and charged hadron irradiated detectors at equal NIEL is an evidence of NIEL hypothesis violation. The diffusion oxygenated float zone detectors are used for the inner-most tracking detectors at LHC, where significant reduction of V_{fd} is required as shown in Fig. 21.15.

The oxygen concentration in DOFZ detectors is around $2 \cdot 10^{17}\,\mathrm{cm}^{-3}$, which is up to an order of magnitude lower than the oxygen concentration in Czochralski (Cz) silicon. They have only recently become available as detector grade material with resistivity (>1 kΩcm) high enough to allow production of 300 μm thick detectors [45]. The increase of V_{fd} after irradiation was found to be smaller or equal to that of DOFZ detectors as shown in Fig. 21.16a. Moreover, for n-type Cz detectors (less evident in p-type Cz) stable donors ($g_c \sim -5 \cdot 10^{-3}\,\mathrm{cm}^{-1}$) are introduced instead of acceptors after fast charged hadron and γ-ray irradiations. The oxygen in form

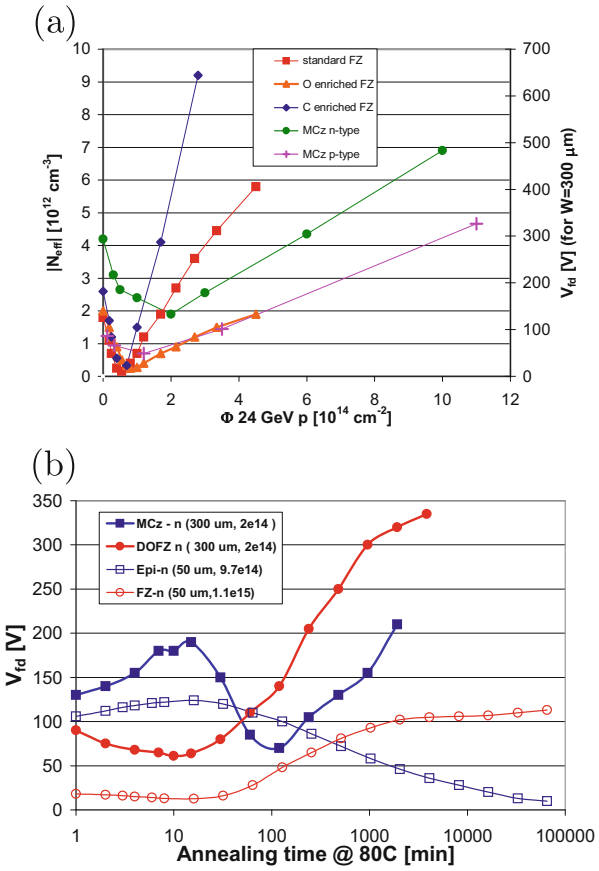


Fig. 21.16 (**a**) Influence of carbon and oxygen enrichment and wafer growth on the change of N_{eff} as function of fluence. (**b**) Annealing of the Magnetic Cz-*n* type (MCz) and diffusion oxygenated samples after $2 \cdot 10^{14}\,\mathrm{cm}^{-2}$. Also shown are thin epitaxial and standard FZ detectors irradiated to fluences around $10^{15}\,\mathrm{cm}^{-2}$. Note the typical behavior of detectors with positive space charge for epitaxial and MCz detectors

of a dimer $[O_{2i}]$, which is more abundant in Cz than FZ detectors, is likely to be responsible. It is a precursor for formation of radiation induced shallow donors (thermal donors) [46]. The reverse annealing in Cz detectors has approximately the same amplitude as in FZ but is delayed to such extent that may not even play an important role at future experiments. The different sign of g_c and g_Y produce a different shape of N_{eff} annealing curve (see Fig. 21.16). During the short term annealing the V_{fd} increases and then starts to decrease as acceptors formed during late stages of annealing compensate the stable donors. Eventually the acceptors prevail and the V_{fd} starts to increase again.

Another interesting material is epitaxial silicon grown on low resistivity Cz substrate [47]. Stable donors are introduced after charge hadron irradiation with

rates depending on the thickness of the epitaxial layer ($g_c = -4 \cdot 10^{-3}$to $-2 \cdot 10^{-2}\,\mathrm{cm}^{-1}$, for thickness of 150–25 μm). They exhibit also the smallest increase of $|N_{eff}|$ after neutron irradiations, but are only available in thicknesses up to 150 μm.

Control of Space Charge

The opposite sign of g_c and g_Y and $|g_Y| > |g_c|$ opens a possibility to control V_{fd} with a proper operation scenario and to keep it low enough to assure good charge collection (see Fig. 21.16b).

This has been demonstrated with thin epitaxial detectors which were irradiated in steps to $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$ and annealed for 50 min at 80°C during the steps which is roughly equivalent to room temperature storage during non-operation periods at LHC or HL-LHC (see Fig. 21.17) [48]. The compensation of stable donors by acceptors activated during the irradiation steps resulted in lower V_{fd} after $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$ than the initial V_{fd}. Allowing detectors to anneal at room temperature during non-operation periods has also a beneficial effect on leakage current and trapping probability as will be shown later.

The use of silicon material with opposite sign of stable damage for neutrons and charged hadrons can be beneficial in radiation fields with both neutron and charged hadron content. The stable acceptors introduced by neutron irradiation compensate stable donors from charged hadron irradiations and lead to reduction of V_{fd} as demonstrated in [49]. An example is shown in Fig. 21.17b for MCz n-type pad detectors which were irradiated by 23 GeV protons (open symbols) and then by neutrons (solid symbols). The additional neutron irradiation decreases the V_{fd}.

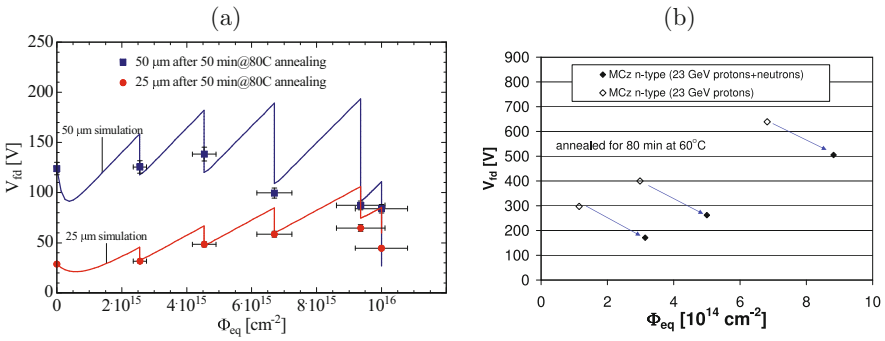


Fig. 21.17 (**a**) An example of space charge compensation through annealing in a thin epitaxial detector irradiated with 23 GeV protons to $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$. The lines denote the Hamburg model prediction. (**b**) Beneficial effect of irradiations by protons and neutrons on V_{fd} for MCz n-type detectors

21.4.2.2 Electric Field

The occupation probability (Eq. (21.15)) of a deep level is determined by its position in the band gap, temperature and concentration of free carriers. The occupancy of initial shallow dopants is largely unaffected by p, n, T and N_{eff} is constant over the entire bulk. The irradiation introduces deep levels which act as generation centers. Thermally generated carriers drift in the electric field to opposite sides (bulk generation current). The concentration of holes is thus larger at the p^+ contact and of electrons at the n^+ contact. Some of these carriers are trapped and alter the space charge i.e. steady state P_t in Eq. (21.14). As a result the N_{eff} is no longer uniform, but shows a spatial dependence, with more positive space charge at p^+ and more negative at n^+ contact Such a space charge distribution leads to an electric field profile different from linear.

The electric field profile can be probed by measuring the current induced by the motion of carriers generated close to an electrode (so called Transient Current Technique). They drift over the entire thickness of detector. The measured induced current at time t after the injection, is then proportional to the electric field, at the position of the drifting charge at time t according to equation $i = -q\vec{E}_w \cdot \vec{v}$. An example of such a measurement can be seen in Fig. 21.18b, where carriers at the back of the detector (n^+ contact) are generated close to electrode by a short pulse of red light. The shape of the current depends on the voltage and temperature. At lower voltages and higher temperatures the electric field shows two peaks, which can only be explained by the space charge of different signs at both contacts. This is usually referred to as "double junction" profile [50, 51], the name indicating that the profile is such as if there were two different junctions at both contacts ($p^+ - n\ p - n^+$ structure). This is evident for under-depleted detectors where both junctions are separated by an un-depleted bulk. Usually one of the regions dominates spatially

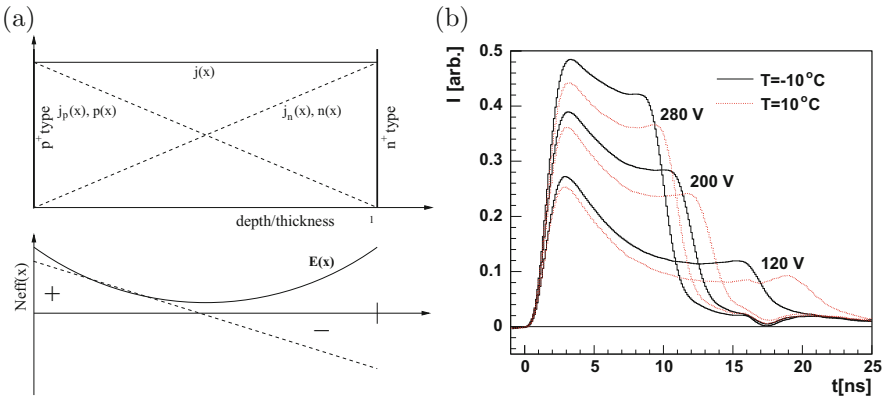


Fig. 21.18 (**a**) Illustration of mechanism leading to non-uniform N_{eff}. (**b**) Induced current due to drift of holes from n^+ side to p^+ side in 300 μm thick oxygenated detector irradiated with 23 GeV protons to $2 \cdot 10^{14}$ cm^{-2}

(also called the "main junction") which determines the predominant sign of the space charge and annealing properties. The space charge profile depends on the balance between the deep levels which occupation depends on n, p and shallow defects mostly unaffected by n, p.

Apart from thermally generated carriers the non-equilibrium carriers which modify the electric field can also be generated by ionizing particles or continuous illumination of detector by light [52].

Modeling of the Field

Even more precise insight in electric field, particularly for heavily irradiated detector ($\Phi_{eq} > 10^{15}\,\mathrm{cm}^{-2}$), is obtained by a more elaborate technique called Edge-TCT [53] shown in Fig. 21.19, where the polished edge of the silicon strip detector is illuminated by narrow beam of infra-red light. The induced current measured promptly after light injection is proportional to the sum of the drift velocities of electrons and holes at a given depth of injection. The drift velocity profile of an detector is hence obtained by scanning over the edge of the detector at different depths. The profiles of heavily irradiated silicon detectors are shown in Fig. 21.20.

The velocity profile in detector moderately irradiated with neutrons (Fig. 21.20a) deviates only slightly from simple model of constant N_{eff} inside the bulk, while at higher fluence (Fig. 21.20b) the electric field shows typical "double junction" behavior, with some remarkable features:

- the main junction penetrates deeper than expected using g_c measured at low fluences
- the high field region at the back extends deep into the detector
- the electric field is present in the whole bulk even at very modest voltages

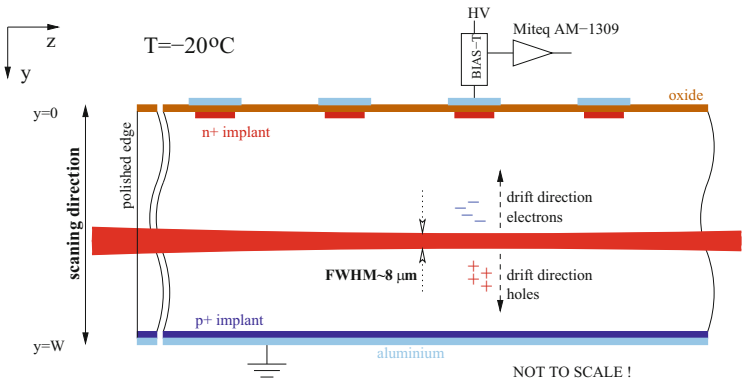


Fig. 21.19 The principle of the Edge-TCT technique

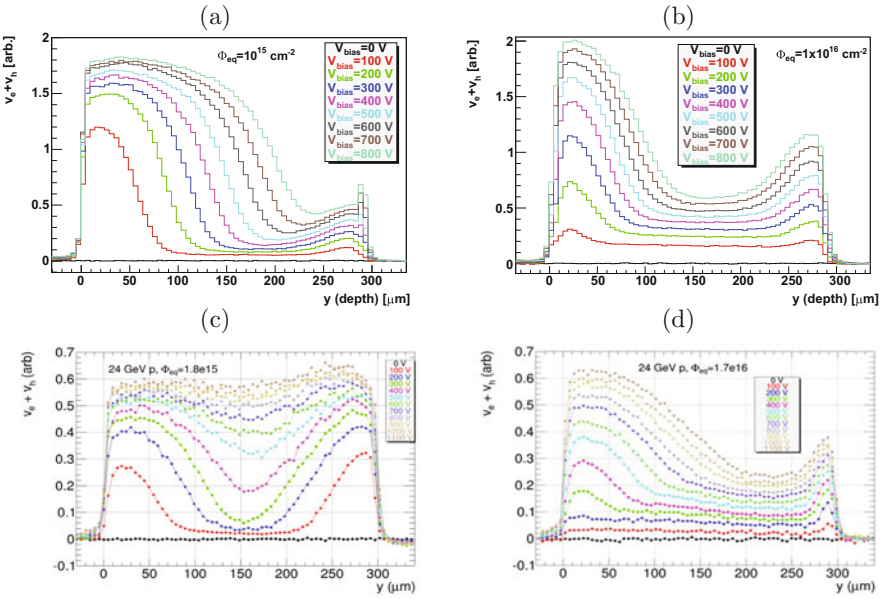


Fig. 21.20 The velocity profiles of neutron irradiated detectors to (**a**) $\Phi_{eq} = 10^{15}\,\mathrm{cm}^{-2}$, (**b**) $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$ and 23 GeV proton irradiates detectors to (**c**) $\Phi_{eq} = 1.8 \cdot 10^{15}\,\mathrm{cm}^{-2}$ and (**d**) $\Phi_{eq} = 1.7 \cdot 10^{16}\,\mathrm{cm}^{-2}$. The measurements were performed with 300 μm thick ATLAS-07 prototype strip detectors with 100 μm pitch and 20 μm implant width at −20 °C. Strips are at $y = 0$ μm

- the velocity in the neutral bulk is very high reaching almost a third of the saturation velocity at very high bias voltages

The appearance of the electric field in the neutral bulk can be explained by the increase of undepleted bulk resistivity and increase of generation current. As both increase also higher field is required for transport of thermally generated carriers across the detector in a steady state.

The electric field in charged hadron irradiated detectors is almost symmetrical at lower fluence (Fig. 21.20c) and becomes similar to neutron irradiated ones only at very high bias voltages (Fig. 21.20d). Already at 500 V the detector is fully active after receiving $\Phi_{eq} = 1.8 \cdot 10^{15}\,\mathrm{cm}^{-2}$. The reason for such behavior is not clear, but points to higher oxygen content of the silicon wafers and different energy levels associated with changes of N_{eff} with respect to the neutron irradiated detectors.

Extraction of electric field from velocity profile is not straightforward [53], due to large uncertainties arising from saturation of drift velocity with the electric field. Instead of precisely modeling $N_{eff}(y)$ several key parameters can be extracted from the measured velocity profiles which can be used to constrain/anchor any electric field model, either effective or calculated from known defects. These parameters are

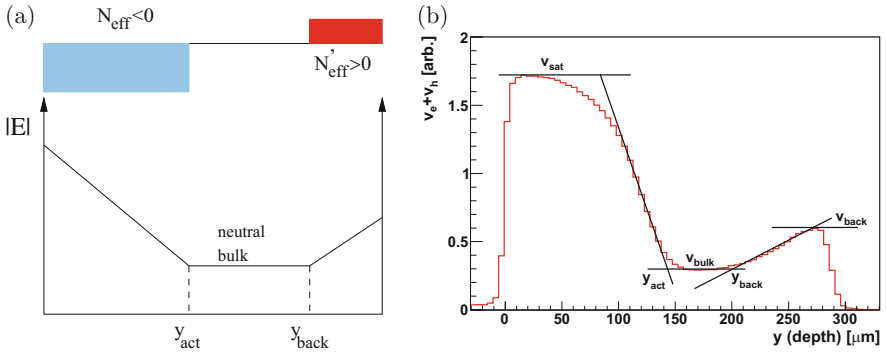


Fig. 21.21 (**a**) Simplest effective space charge and electric field model in irradiated strip detectors. (**b**) Extraction of key parameters determining electric field from the measured velocity profile

shown in Fig. 21.21 and are:

- depth of active region with negative space charge extending from the electrode side y_{act}
- velocity in undepleted bulk v_{bulk}
- depth of positive space charge region at the back of the detector $W - y_{back}$
- velocity at the back of the detector v_{back}

The parameters extracted for neutron irradiated detectors are shown in Fig. 21.22. The change of active region depth y_{act} with voltage is compatible with g_c up to the fluence of $\Phi_{eq} < 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$, while a three times lower g_c was extracted at $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$. Drift velocity in neutral bulk increases both with fluence and voltage, while the depth of the active region at the back is less dependent on fluence.

It is clear that in heavily irradiated detectors ($\Phi_{eq} > 1 - 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$) the V_{fd} doesn't serve as a relevant parameter determining the active thickness as the whole detector becomes active with irradiation.

21.4.2.3 Charge Multiplication

The increase of N_{eff} with irradiation and high applied bias voltages lead to very high electric fields close to electrodes. They can become high enough so that the electrons gain enough energy in its free path to create new e–h pairs, a process called impact ionization. After drifting over the distance dx the number of free carriers increases by

$$dN_{e,h} = \alpha_{e,h}\, N_{e,h}\, dx \tag{21.34}$$

where $\alpha_{e,h}$ are the impact ionization coefficients for electrons and holes [54, 55]. Charge multiplication through impact ionization is a well known process and widely

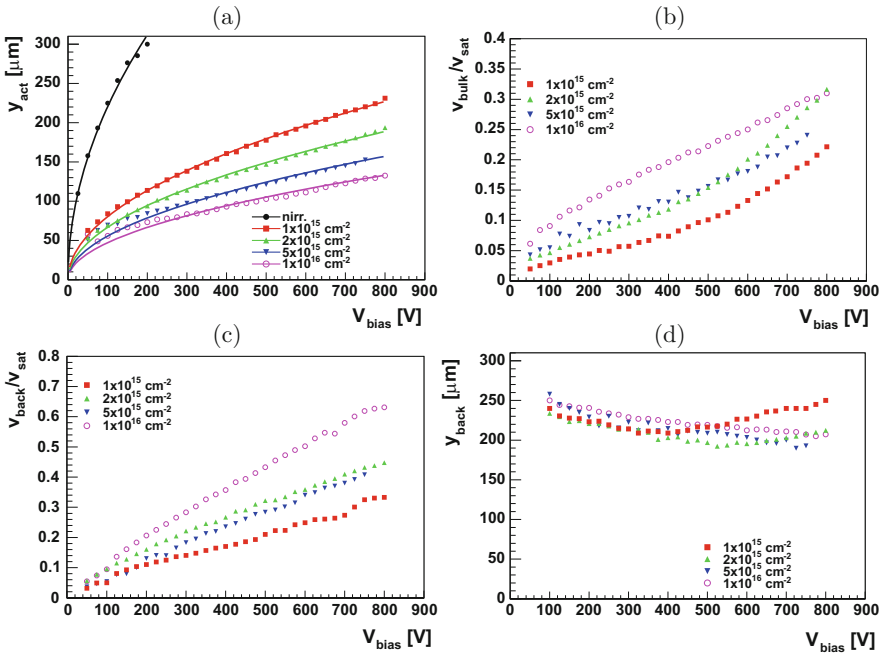


Fig. 21.22 The relevant parameters of the electric field in the neutron irradiated silicon detector—see Fig. 21.21 for explanation

exploited in Avalanche Photo Diodes and Si-Photo-multipliers. It was however not observed directly in irradiated silicon detectors. Prediction of detector performance a decade ago based on extrapolation of damage parameters to fluences well above $\Phi_{eq} > 10^{15}\,\mathrm{cm}^{-2}$ greatly underestimated the charge collection and detection efficiency.

Part of this, better than expected, performance can be attributed to favorable electric field profile, part to smaller trapping (discussed later) and part to charge multiplication. A key factor was improved high voltage tolerance of detectors which allowed application of bias voltages exceeding 1 kV.

Charge multiplication has since been undoubtedly observed with charge collection efficiency $CCE > 1$ in pad detectors [56], 3D detectors [57] and mostly strip detectors [58, 59] (see Fig. 21.23a). Another direct evidence came from TCT measurements where the drift of holes produced in multiplication was clearly observed as shown in Fig. 21.23b. There are several aspects of charge multiplication that make it difficult to control and master:

- Charge multiplication is geometry/process dependent; fields between 15–25 V/μm are required to produce sizable gain (∼1 e_0/μm). To achieve high gains the shape of implant and segmentation of electrodes (pitch and implant width) are very important. Strong field focusing close to implant edges leads to

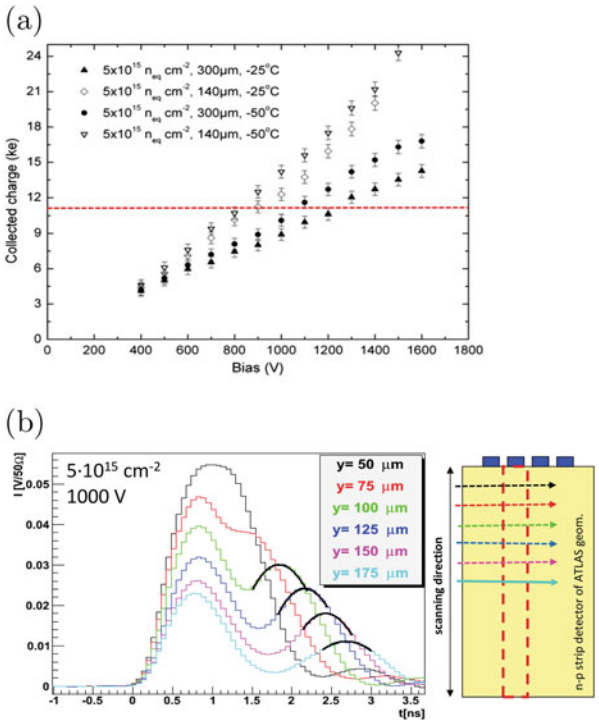


Fig. 21.23 (**a**) Measured charge collection dependence on voltage for 140 and 300 μm thick strip detectors. The red line denotes the charge measured in non-irradiated 140 μm thick detector. (**b**) Induced current pulses in strip detector for different depths of Edge-TCT injection. The second peak in the induced current pulses is due to multiplied holes drift

higher gains. This is also the reasons why larger gains were observed in highly segmented detectors.

- Charge gain depends on the hit position within the electrode. In highly irradiated strip detectors higher gain was observed for tracks few μm away from the implant, where the electric fields are highest [60].
- The holes produced in multiplication are trapped by deep defects (change of free hole concentration, p, in Eq. (21.15)) which reduce the negative space charge—act as a feedback. Therefore gain increases moderately with voltage and is usually limited to factors below <10.
- Gain can vary on time scale of days when detector is under bias [61].
- It is difficult to parametrize the field and reliably simulate the operation.

Annealing Performance of Highly Irradiated *p*-type Detectors Active bulk and charge multiplication have an important impact on performance of p-type detectors after annealing. Increase of N_{eff} with time and consequent increase of electric field increases gain. On the other hand smaller high field region near the electrodes affects less the performance due to significant field in the neutral bulk. A typical annealing

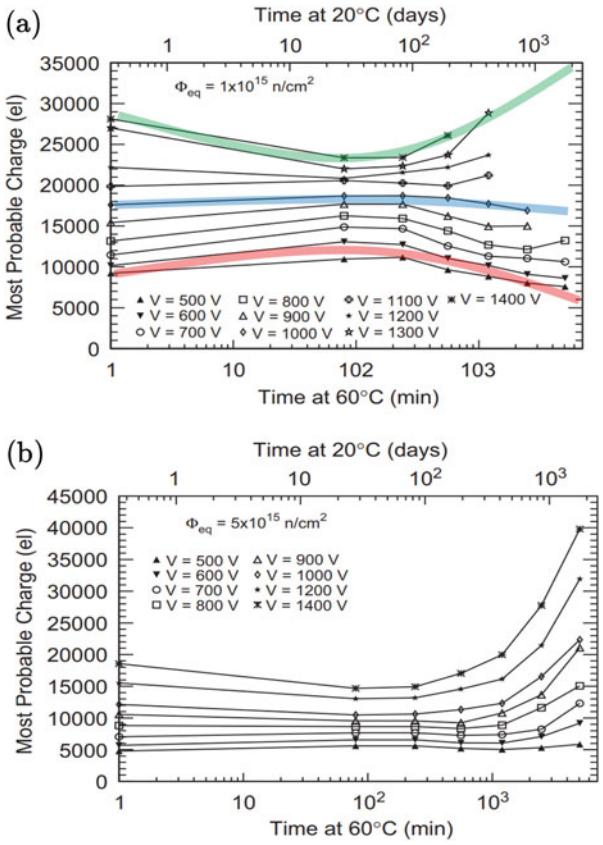


Fig. 21.24 Dependence of charge collection on annealing time at 60 °C at different bias voltages at (**a**) $\Phi_{eq} = 1 \cdot 10^{15}$ cm^{-2} and (**b**) $\Phi_{eq} = 5 \cdot 10^{15}$ cm^{-2} [62]

performance is shown in Fig. 21.24a. At lower voltages charge collection increases during short term and decreases during long term annealing (red band), which is in agreement with evolution of effective doping concentration. At higher voltages the charge multiplication compensates the decrease of active region (blue band) and at highest voltages overcompensates it, resulting in smallest charge collection for completed short term annealing (green band). At higher fluences and voltages shown in Fig. 21.24b the beneficial effect of long term annealing is even more pronounced.

Noise The increase of noise due to multiplication can diminish the benefits or even deteriorate the performance in terms of signal/noise ratio. The details about the noise in multiplication mode will be discussed at in the section on electronics.

21.4.2.4 Charge Trapping

The decrease of charge collection efficiency is determined by the trapping term and the product $\vec{E} \cdot \vec{E_w}$ in Eq. (21.29). At fluences beyond that at LHC the trapping term dominates and ultimately sets the limit of efficient operation. The influence of trapping on charge collection can be clearly seen for a fully depleted detector, where the degradation of the induced charge is exclusively due to trapping. The collected charge degrades with fluence as shown in Fig. 21.25a. The degradation is severe and around half the charge in non-irradiated detector (12000 e_0) are measured at V_{fd} for $\Phi_{eq} \sim 10^{15}\,\mathrm{cm}^{-2}$. The induced charge increases further for bias voltages larger than V_{fd}. Higher electric field reduces the drift time and by that the influence of trapping term.

If the deep levels responsible for trapping are constant in time or change with a first order process (see Eq. (21.5)), then at any time after irradiation their concentration is linearly proportional to the fluence. Under this assumption Eq. (21.19) can be rewritten as

$$\frac{1}{\tau_{eff_{e,h}}} = \frac{1}{\tau_{eff0_{e,h}}} + \beta_{e,h}(t,T)\,\Phi_{\mathrm{eq}}, \tag{21.35}$$

where $\beta_{e,h}$ is called effective electron and hole trapping damage constant which depends on temperature, time after irradiation and irradiation particle. In detector grade silicon the effective trapping probability of a non-irradiated detector $\frac{1}{\tau_{eff0_{e,h}}}$ is negligible and is usually omitted from Eq. (21.35). Alternatively the trapping distance can be defined as

$$\lambda_{e,h} = \mu_{e,h}\,\tau_{eff_{e,h}}\,E \tag{21.36}$$

measuring the distance the carriers drift before being trapped.

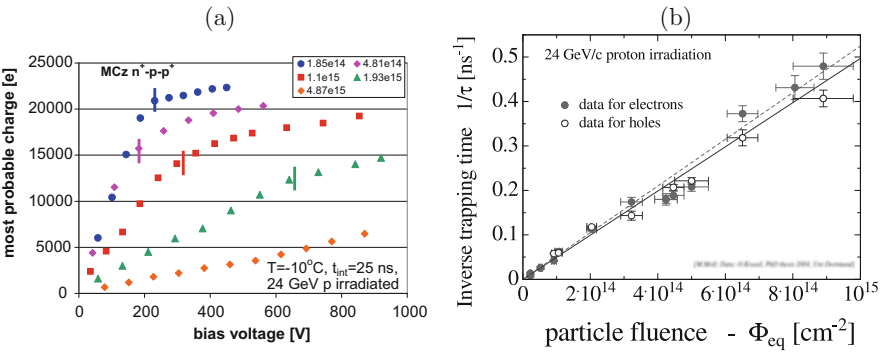


Fig. 21.25 (**a**) Dependence of induced charge on voltage for MCz p-type pad detector irradiated to different fluences. The V_{fd} for each measurement is denoted by vertical bar. (**b**) Effective trapping times of electrons and holes as found in Ref. [64]

The trapping times in silicon were systematically measured with Transient Current Technique [63]. The trapping probabilities for 23 GeV protons are shown in Fig. 21.25b. At $\Phi_{eq} \sim 10^{15}\,\mathrm{cm}^{-2}$ the effective trapping times are around few ns.

The trapping damage constant was studied as a function of different material properties: resistivity, oxygen concentration, carbon concentration, wafer production (MCz, FZ, epi-Si) and type of silicon (p-type or n-type). It was found, within the error margin, not to depend on any, thus being universal for silicon. The average values of β for neutrons and charged hadrons are given in the Table 21.5 [65]. It shows that the trapping probability for electrons is smaller than for holes. The NIEL hypothesis is slightly violated as charged hadrons produce more damage than reactor neutrons.

The evolution of trapping probability with time after irradiation is described in the simplest model by the decay of the dominant trap to another dominant trap (Eq. (21.3)) or a model with two traps one constant in time and one that decays. Both models can be described by the following equation [63]

$$\beta_{e,h}(t) = \beta_{0_{e,h}} \cdot e^{-\frac{t}{\tau_{ta,e,h}}} + \beta_{\infty_{e,h}} \cdot (1 - e^{-\frac{t}{\tau_{ta,e,h}}}) \tag{21.37}$$

with $\beta_{0_{e,h}}$ and $\beta_{\infty_{e,h}}$ the trapping rates at early and late annealing times, respectively. For the annealing temperatures of interest β_0 is very close to β measured at the end of short term annealing ($\beta(t_{min})$) given in Table 21.5. There is a distinctive difference between annealing of effective trapping times for holes and electrons. The trapping probability of holes increases with annealing time and that of electrons decreases (see Fig. 21.26) irrespective of material properties and type of irradiation

Table 21.5 Trapping time damage constants for neutron and fast charged hadron irradiated silicon detectors measured after the end of short term annealing [65]

t_{min}, $T = -10°\mathrm{C}$	β_h [$10^{-16}\,\mathrm{cm}^{-2}$/ns]	β_e [$10^{-16}\,\mathrm{cm}^{-2}$/ns]
Reactor neutrons	4.7 ± 1.2	3.5 ± 0.6
Fast charged hadrons	6.6 ± 1.1	5.3 ± 0.5

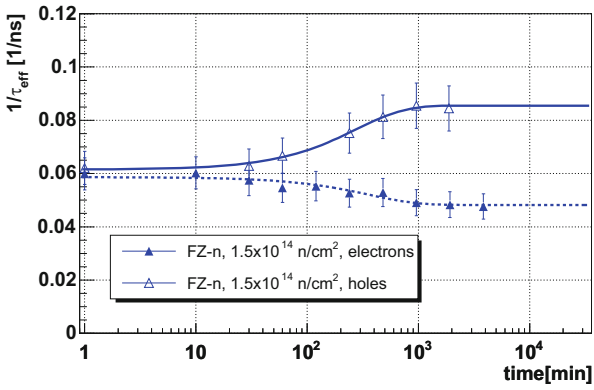


Fig. 21.26 Annealing of $1/\tau_{eff,e,h}$ for a detector irradiated with neutrons to $\Phi_{eq} = 1.5 \cdot 10^{14}\,\mathrm{cm}^{-2}$

Table 21.6 Parameters used to model annealing of effective trapping times

	τ_{ta} [min at 60 °C]	$(\beta_0 - \beta_\infty)/\beta_0)$	E_{ta} [eV]
Electrons	650±250	0.35±0.15	1.06 ± 0.1
Holes	530±250	0.4±0.2	0.98 ± 0.1

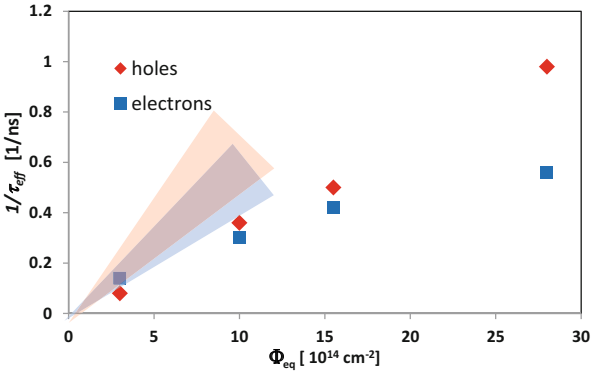


Fig. 21.27 Effective trapping probability measured at high fluences of charged hadrons [66]. The red and blue bands indicate the predictions of trapping probability of holes and electrons from Table 21.5

particle. The parameters describing annealing of effective trapping probabilities are shown in Table 21.6. The activation energy E_{ta} should be used in Eq. (21.9) for scaling τ_{ta} to different temperatures. The $\beta_{e,h}$ depends only moderately on temperature [63]. At temperatures of interest for most applications the trapping probabilities for both holes and electrons decrease with temperature by around 10–20% if the temperature changes from −20° to 20 °C.

The linear relation of Eq. (21.35) breaks down at equivalent fluences higher than $\sim 10^{15}\,\mathrm{cm}^{-2}$, where it starts to exhibit saturation. Unfortunately the TCT can not be directly used to measure trapping probabilities and values have be extracted by combining both TCT and CCE measurements with simulations. The study performed by CMS collaboration is shown in Fig. 21.27 [66]. It can be seen that already at few times $10^{15}\,\mathrm{cm}^{-2}$ the effective trapping probabilities deviate significantly from linear. Recently studies [67] showed that at extreme fluences of $\sim 10^{17}\,\mathrm{cm}^{-2}$ the trapping probability is around an order of magnitude smaller than predicted from the low fluence measurements.

21.4.2.5 Generation Current

The defects influencing the generation current (Eq. (21.22)) were found to either dissociate or are constant in time. The bulk damage-induced increase of the reverse current (ΔI) exhibits therefore a simple dependence on particle equivalent fluence at any time after irradiation

$$\Delta I_{gen} = \alpha(t, T)\, V\, \Phi_{eq}, \tag{21.38}$$

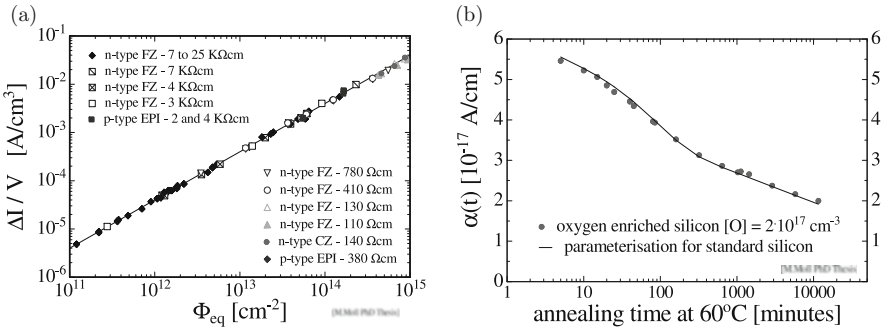


Fig. 21.28 (**a**) Dependence of bulk generation current on fluence for different detectors after 80 min storage at 60 °C. (**b**) Annealing of leakage current damage constant (after [31])

where V is the active volume ($V = S\,w$) and α the leakage current damage constant. The bulk generation current scales with NIEL, hence the leakage current damage constant is independent of the silicon properties and irradiation particle type as shown in Fig. 21.28a [68]. The measured value of the leakage current depends exponentially on the operating temperature as (see terms in Eq. (21.22))

$$I_{gen}(T) \propto T^2 \exp\left(-E_g/2k_BT\right), \tag{21.39}$$

and accordingly all α-values can be scaled to any temperature.

The damage induced bulk current undergoes also a temperature dependent beneficial annealing, described by

$$\alpha(t) = \alpha_1 \exp\left(-\frac{t}{\tau_\alpha}\right) + \alpha_0 - \alpha_2 \ln\left(\frac{t}{t_{norm}}\right), \tag{21.40}$$

with $\alpha_0 = 5.03 \cdot 10^{-17}$ A/cm, $\alpha_1 = 1.01 \cdot 10^{-17}$ A/cm, $\alpha_2 = 3.34 \cdot 10^{-18}$ A/cm, $\tau_\alpha =$ 93 min and $t_{norm} = 1$ min all measured at 60 °C. The first term in the Eq. (21.40) describes the decay of the defect and the second contribution of the defects constant in time. The last term is associated with the decay of the cluster, a conclusion based on its absence in ^{60}Co irradiations [68]. The leakage current annealing can be seen in Fig. 21.28b. Universality of the annealing described by Eq. (21.40) can be used to reliably monitor the equivalent fluence of particle sources even in cases of wide energy distributions. As a standard α(80 min at 60°C, 20°C) $= 4 \cdot 10^{-17}$ A cm^{-1} is used.

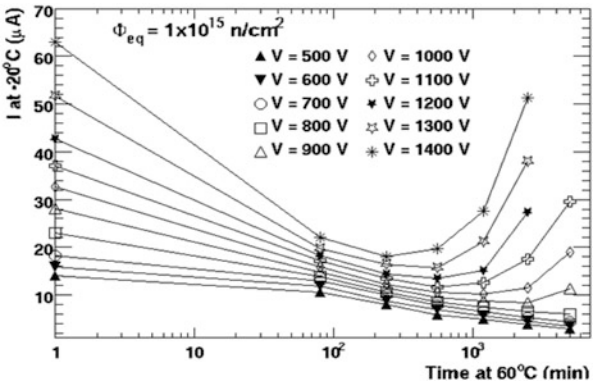


Fig. 21.29 Dependence of leakage current on annealing time at different voltages. The increase of leakage current with annealing is due to charge multiplication (from Ref. [62])

Leakage Current in Presence of Charge Multiplication

For devices with gain the leakage current is given by the current gain M[2] and generation current $I = M \cdot I_{gen}$. An example of the leakage current increase at high bias voltages during annealing is shown in Fig. 21.29. One should however be careful as the increase of leakage current at high bias voltages can also be attributed to other effects such as the onset of thermal runaway or rise of the surface current, however without clear increase of the collected charge.

21.4.2.6 Alternative Ways of Operation

The key reason for changes in performance of an irradiated detector are deep traps. The manipulation of their occupancy therefore has an influence on the detector properties. Variation of the operation temperature and/or concentration of free carriers can be used to change the occupancy of deep traps. The first observation of charge collection efficiency recovery after gradually cooling down the heavily irradiated silicon detector from room temperature to cryogenic temperatures (see Fig. 21.30a) was reported in [69] and referred to as "Lazarus effect". However the operation of silicon detectors under reverse bias turned out to be very sensitive to previous biasing conditions and ionizing particle rates. The signal varies with time after exposure to ionizing particles as shown in Fig. 21.30b. The trapping of the drifting carriers enhances the space charge of different signs at both detector contacts (see Sect. 21.4.2.2) to the point where the applied voltage is insufficient to establish the electric field in the entire detector. As a consequence the charge

[2]Current and charge gains can be in principle different, but have been so far observed to be very similar.

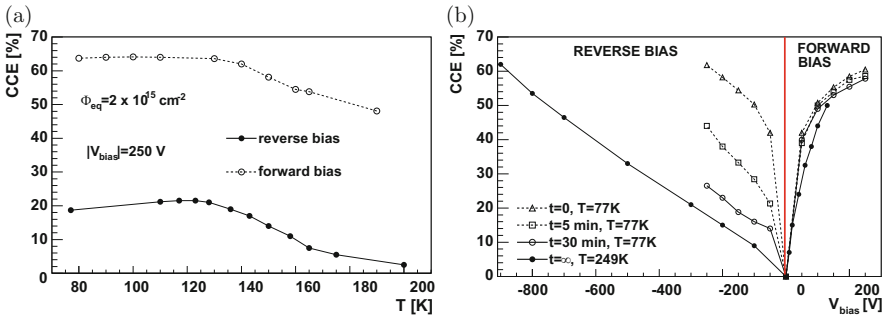


Fig. 21.30 (**a**) Charge collection efficiency in 400 μm thick detector irradiated to 10^{15} cm^{-2} in forward and reverse direction. (**b**) The dependence of CCE on voltage at $T = 77$ K in both forward and reverse direction of a detector irradiated to $2 \cdot 10^{15}$ cm^{-2}

collection efficiency is reduced. The phenomena of polarization of the detector by trapped charge is not unique to silicon and is present also in other semiconductors. Since emission times depend on $E_g/(2\,k_B\,T)$, silicon at cryogenic temperatures behaves similarly as wide band gap semiconductors at room temperature.

At cryogenic temperatures a more stable operation is achieved with detectors biased in forward direction [70] (see Fig. 21.30a,b). The resistivity of the bulk increases with irradiation and it effectively becomes a heavily doped insulator. Applied bias in forward direction injects carriers in the detector. These are trapped at deep levels and affect the electric field. The predominately negative space charge is naturally compensated by injection of holes. The electric field grows from $E \approx 0$ at the injection point towards the other contact with the square root of the distance x from the injecting junction [71]

$$E(x) = \frac{3\,V}{2\,W}\sqrt{\frac{x}{W}}. \tag{21.41}$$

The electric field extends through the entire detector thickness regardless of the applied voltage or concentration of the deep levels. This is an important advantage over the biasing of detectors in reverse polarity. The drawback of forward bias operation is the increased current, requiring intensive cooling. The current dependence on voltage is quadratic ($I \propto V^2$), followed by a sharp rise at threshold voltage V_T as shown in Fig. 21.31. It happens when the space charge saturates due to filling all the traps and current can not be limited by increasing the concentration of the trapped carriers, therefore $V_T \propto \Phi_{eq}$. An important feature of this mode of operation is the fact that the current at a given voltage progressively decreases with fluence (see Fig. 21.31), approximately as $I(\Phi_{eq}) \propto \Phi_{\mathrm{eq}}^{-1.5}$. The larger the concentration of traps the smaller is the current which is needed to adjust the electric field. Nevertheless, it is still larger than in reverse direction.

In principle, a $p^+ - n - n^+$ structure should inject holes and electrons, which would not produce the aforementioned properties. However it turns out that at n^+

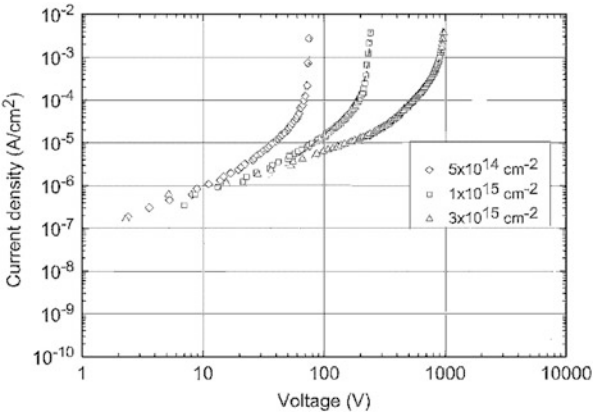


Fig. 21.31 Leakage current-voltage characteristics in forward mode of operation

contacts electrons are not injected [71]. The symmetric structure $p^+ - n - p^+$, where only holes are injected, has the same properties pointing to the same underlying physics process. The same condition of carrier injection can be also achieved in reverse bias mode by continuous illumination of one side by light of short penetration depth [72]. The injected carries establish the same condition as under forward bias and the Eq. (21.41) applies.

The filling of deep levels affects effective trapping probabilities of electrons and holes. Measurements have shown that the same charge collection efficiency is achieved as for a fully depleted detector at few times smaller bias voltage [70, 73] (see Fig. 21.30b). Smaller bias results in smaller average electric field and therefore longer collection times. As the reduction of charge collection efficiency depends in first approximation on the ratio of the drift time to the trapping time of the carriers, the latter must be longer than under the reverse bias.

It is obvious that forward bias operation mode becomes usable once the detectors are already heavily irradiated. There are two ways of how to use detectors in real experiments. With read-out electronics sensitive to both polarities detectors can be first used in reverse and later in forward direction or the detectors are irradiated before being used. In general the use of the forward bias means replacing the problem of the high voltage required for the reverse bias operation by the problem of a high dark current. Therefore detectors with small element size (i.e. pixels) are more suitable for this mode of operation.

21.4.2.7 3D Detectors—A Radiation Harder Detector Design

One approach to address the issue of radiation damage are optimized detector geometries. A good example of radiation hard detector design are so called 3D detectors. An schematic view of such detector is shown in Fig. 21.32 [74]. The

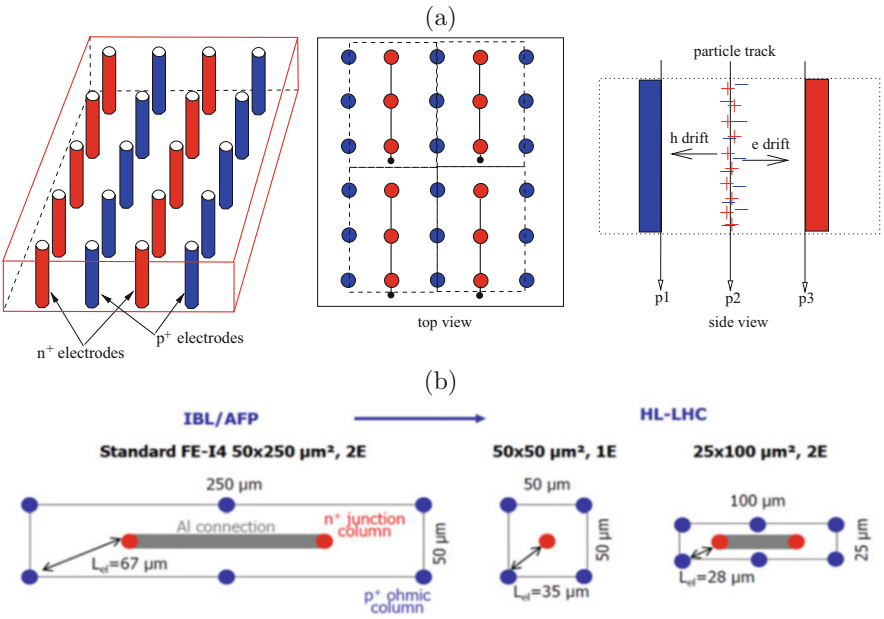


Fig. 21.32 (**a**) The schematic view of the 3D detector (left). The view of the detector surface (middle); gray n^+ electrodes, dark gray p^+ electrodes, black metal line, black dot the bump-bond. The dashed line marks the pixel cell with three columns. Q_e^t/Q^t for different tracks: p1=1, p2=0.5 in p3=0 (right). (**b**) Layout of a single cell/pixel of an IBL 3D detector (2 electrode configuration—2E) and of HL-LHC detector with both options 1E and 2E. The maximum drift length of carriers is indicated

electrodes in such detectors are perpendicular to the surface. Such placement of electrodes has two beneficiary effects for heavily irradiated detectors. The small distance between the electrodes effectively reduces the full depletion voltage. Even more importantly, the drift length of carriers is reduced and therefore the probability of drifting carriers to get trapped ($\tau_{eff,e,h} \gg t_{drift}$). As the signal (number of e-h pairs in Eq. (21.29)) is determined by the detector thickness, vertical electrode configuration ensures good charge collection at moderate voltages. Several columns can be connected together to form pixel cells or strips (Fig. 21.32b). The thickness of the detector is limited by the deep reactive ion etching process used to produce holes. The standard aspect ratio (hole length/hole diameter) is around 24. Apart from a more complex processing, which can be simplified by electrodes not penetrating fully the detector [75, 76], there are some drawbacks of the 3D design:

- Reducing the inter-column spacing results in higher inter-electrode capacitance
- Columnar electrodes are a non-active part of the detector volume and can lead to particle detection inefficiency; most of the tracks in experiments are, however, inclined which mitigates the problem.

- Induced charge depends on the hit position of the particle track. Unlike in planar detectors, the ratio Q_e^t/Q^t varies between 0 and 1 across the detector and can affect the position resolution and efficiency (see Fig. 21.32c).

Nevertheless, these detectors are often first choice for tracking detectors at highest fluences. ATLAS pixel detector (Insertable B Layer—IBL) [77] saw the first application of 3D detectors for tracking in high energy experiments, covering 25% of the total IBL surface at both sides of the staves. The 3D technology is improving with different ways of processing the detectors with single-sided process or more elaborate double sided processing with possibility of active/slim edges reducing the inactive part at the detector border. Efficient charge collection was achieved also for sensors where columns don't penetrate the whole depth. Such a design improves the yield of sensor production. The latter remains one of the main concerns for 3D technology reaching around 50–60% for the IBL module production [78].

At HL-LHC the 3D detectors are planned for the first pixel layer. A small cell size will have a single junction column (cell $50 \times 50\,\mu m^2$) or two columns (cell $25 \times 100\,\mu m^2$), where the maximum drift distances will be reduced to mere 37 and 28 μm making these detectors extremely radiation hard. The first beam tests with such 230 μm thick detectors showed [79] 97% detection efficiency for perpendicular tracks after extreme fluences of $2.5 \cdot 10^{16}\,cm^{-2}$ at >200 V using IBL readout electronics (FE-I4) [80].

21.4.2.8 Timing Detectors

At HL-LHC coping with large particle fluxes emerging from collisions will be an enormous challenge. On average 200 p-p collisions will occur every 25 ns, with collision points distributed normally along the beam with $\sigma_z = 5$ cm and in time with $\sigma_t = 180$ ps. Resulting track and jet densities in the detector complicate the analysis of the underlying reactions that took place. A way to cope with that problem is separation of individual collisions also in terms of time of occurrence within each bunch crossing. This is particularly important for tracks/jets in forward direction for which the position resolution of primary vertex is much worse (∼1 mm). If tracks are not resolved in time, this can lead to false vertex merging. A timing resolution of around 30 ps with respect to the HL-LHC clock is required to successfully cope with pileup. Such an outstanding single particle timing resolution was up to recently impossible with silicon detectors.

Three factors determine the timing resolution of each sensor: time walk which is a consequence of non-homogeneous charge deposition by an impinging particle, noise jitter ($\sigma_{jitter} = t_{rise}/(S/N)$) and resolution of time-to-digital conversion. Standard silicon detectors of 300 μm are not appropriate for precise timing measurement as the integration time to collected all the charge and consequent rise time t_{rise} are large, hence the jitter. In addition fluctuations, not only of the amount of the

charge (time walk correctable by e.g. constant fraction discrimination), but also of the deposition pattern (non-correctable time walk)—so called Landau fluctuations—ultimately limit the time resolution to effectively >100 ps [81]. High enough signal-to-noise S/N in thin detectors can be achieved by using so called Low Gain Avalanche Detectors (LGAD) [82].

They are based on a $n^{++} - p^{+} - p - p^{++}$ structure where an appropriate doping of the multiplication layer (p^{+}) leads to high enough electric fields for impact ionization (see Fig. 21.33) [82]. Gain factors in charge of few tens significantly improve the resolution of timing measurements, particularly for thin detectors. The main obstacle for their operation is the decrease of gain with irradiation, attributed to effective acceptor removal in the gain layer [41]. A comprehensive review of time measurements with LGADs is given in Ref. [83].

The most probable charge in 50 μm and 80 μm thick pad devices before and after irradiation is shown in Fig. 21.34a. As soon as multiplication layer is depleted the gain appears. At lower fluences the gain degradation at the depletion of multiplication layer (around 40 V) can be clearly seen. At higher fluences the gain appears at high bias voltages where over-depletion ensures that high enough electrical fields are reached; above $\Phi_{eq} > 10^{15}\,\mathrm{cm}^{-2}$ the onset of multiplication is observed only at highest voltages of around 700 V. Such voltages correspond to very high average fields of 15 V/μm. At fluences $\Phi_{eq} > 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$ the beneficial effect of multiplication layer is gone. The devices of the same design without multiplication layer show similar behavior as LGADs. The time resolution of LGADs was extensively measured in the test beams [84] and with ^{90}Sr electrons. It is shown in Fig. 21.34b for the 50 μm thick non-irradiated devices.

At very large fluences of $\Phi_{eq} > 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$ the gain, although lower than the initial, appears due to deep traps (see section on charge multiplication)

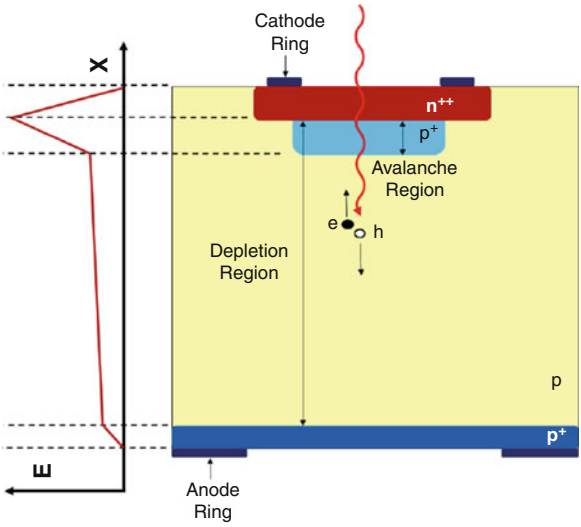


Fig. 21.33 Schematic view of the Low Gain Avalanche Detector

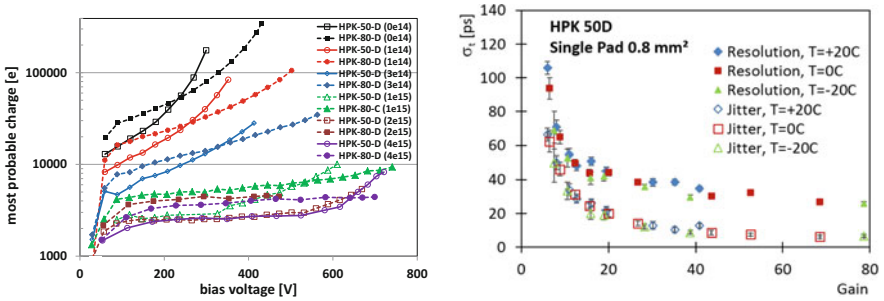


Fig. 21.34 Dependence of most probable charge for irradiated LGAD devices on voltage for different thickness (50 and 80 μm). Fluences in the brackets are in [cm^{-2}] [85]. Around 3000 e is expected for a 50 μm device without gain layer. (**b**) Time resolution and its noise jitter contribution measured for the non-irradiated 50 μm detector at different temperatures [86]

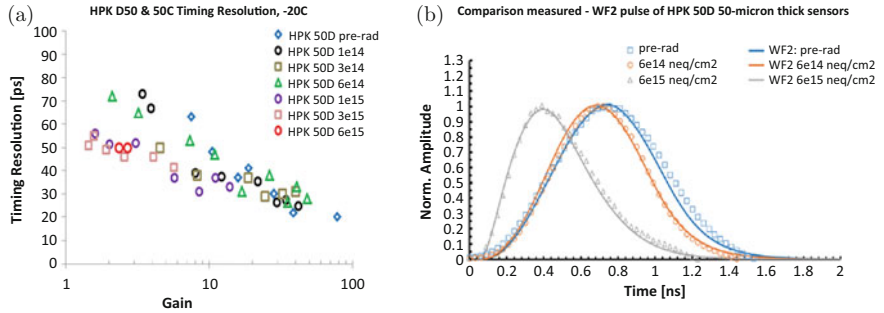


Fig. 21.35 (**a**) Time resolution of irradiated LGAD detectors at different gains and fluences ([cm^{-2}]). (**b**) Measured and simulated induced current pulse shape at different irradiation levels. Note that amplitudes were normalized to one

and the timing resolution degrades only moderately (see Fig. 21.35a). Moreover, multiplication in larger volume of the bulk results in faster rise time of the induced current which at given gain leads to better timing resolution (see Fig. 21.35b).

The leakage current in LGADs follows the same equation as discussed in section on charge multiplication. Hence, the gain can be calculated from measurement of leakage current and calculated generation current [85, 86].

A lot of effort was spend in recent years to increase the radiation hardness of LGADs by mitigating the acceptor removal. The efforts concentrated to use of co-implantation of carbon [87] to multiplication layer aiming to reduce the removal constant or replacing boron with gallium, which should be more difficult to displace [87, 88].

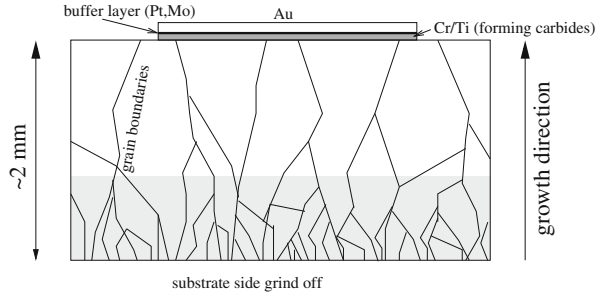


Fig. 21.36 Schematic view of the pCVD diamond detector

21.4.3 Diamond Detectors

Although the specific ionization in diamond detectors is around three times smaller than in silicon, larger detector thickness, small dielectric constant, high break down voltage and negligible leakage current make them the most viable replacement for silicon in the highest radiation fields.

The intrinsic concentration of carriers in diamond is extremely low (good insulator, $\rho > 10^{16}$ Ωcm). The detectors are therefore made from intrinsic diamond metallized at the back and the front (see Fig. 21.36) to form ohmic contacts. Most of the diamond detectors are made from poly-crystalline diamond grown with chemical vapor deposition technique (CVD). Recently also single crystalline (scCVD) detectors have become available. The quality of the poly-crystalline (pCVD) diamond as a particle detector depends on the grain size. The grains in this material are columnar, being smallest on the substrate side, and increasing in size approximately linearly with film thickness.[3] Crystal faults at the boundaries between the grains give rise to states in the band gap acting like trapping centers.

A widely used figure of merit for diamond is its charge collection distance (CCD), which is defined as

$$CCD = \frac{Q_t}{\rho_{e-h}}. \tag{21.42}$$

The CCD represents the average distance over which carriers drift. If $CCD \ll W$, it is equivalent to the trapping distance $\lambda_e + \lambda_h$.[4]. After irradiation, and for pCVD detectors also before irradiation, the CCD depends on electric field, due to reduced probability for charge trapping at larger drift velocity. Only for non-irradiated scCVD detectors $\lambda_{e,h} \rightarrow \infty$ and the $CCD = W$ regardless of the bias voltage

[3]For this reason, many detectors have the substrate side etched or polished away.

[4]The exact relation between the charge collection and the trapping distance is: $CCD = \lambda_e\,[1 - \frac{\lambda_e}{W}(1 - exp(-\frac{W}{\lambda_e}))] + \lambda_h\,[1 - \frac{\lambda_h}{W}(1 - exp(-\frac{W}{\lambda_h}))]$.

applied. Most commonly, the CCD is defined at $E = 1\,\mathrm{V/\mu m}$ or $E = 2\,\mathrm{V/\mu m}$, although sometimes also CCD at saturated drift velocity is stated.

The CCD of typically 500 μm thick diamond detectors has improved tremendously over the last 20 years. Current state of the art pCVD detectors reach up to 300 μm at 2 V/μm and are available from 6 inch wafers.

21.4.3.1 Radiation Hardness

In pCVD detector the leakage current does not increase with irradiation; moreover it may even decrease, which is explained by passivation of defects at grain boundaries. The current density in high quality pCVD diamond is of order 1 pA/cm^2, a value strongly dependent on the quality of metallized contacts.

Irradiation decreases the CCD for both scCVD and pCVD diamonds with similar rate [89], pointing to the in-grain defects being responsible. It has been observed that exposing such an irradiated detector to ionizing radiation (10^{10} minimum ionizing particles/cm^2) improves the charge collection efficiency of pCVD detectors by few 10%. This process is often called "pumping" or priming. The ionizing radiation fills the traps. The occupied traps become inactive, hence the effective trapping probability decreases. The traps can remain occupied for months due to large emission rates if kept in the dark at room temperatures. Once detectors are under bias the ionizing radiation leads to polarization of detectors, in the same way as in silicon, but with the polarization persisting over much longer times. The measurements of charge collection can therefore depend on previous biasing condition and relatively long times are needed to reach steady state of operation.

The irradiation decreases the trapping distance of electrons and holes proportionally to the fluence. The relation can be derived by inserting the effective trapping time (Eq. (21.35)) in the expression for the trapping distance (Eq. (21.36)):

$$\frac{1}{\lambda_{e,h}} = \frac{1}{\lambda_{0_{e,h}}} + K_{e,h} \cdot \Phi, \tag{21.43}$$

where λ_0 denotes the trapping distance of an unirradiated detector and $K_{e,h}$ the damage constant. Assuming $\lambda_e \approx \lambda_h$ and $\lambda_e + \lambda_h \ll W$ for simplicity reasons, CCD dependence on fluence can be calculated as:

$$\frac{1}{CCD} \approx \frac{1}{CCD_0} + K\,\Phi. \tag{21.44}$$

Although only approximate the Eq. (21.44) fits the measurements well over a large fluence range as shown in Fig. 21.37a. The extracted damage constant K ($\sim 1/2\,K_{e,h}$) from source and test beam data for particles of different energy and spectrum are gathered in Table 21.7. For high fluences the second term in Eq. (21.44) prevails and the scCVD and pCVD diamonds perform similarly. At $\Phi = 2 \cdot 10^{16}\,\mathrm{cm}^{-2}$ of 23 GeV protons the $CCD \approx 75$ μm which corresponds

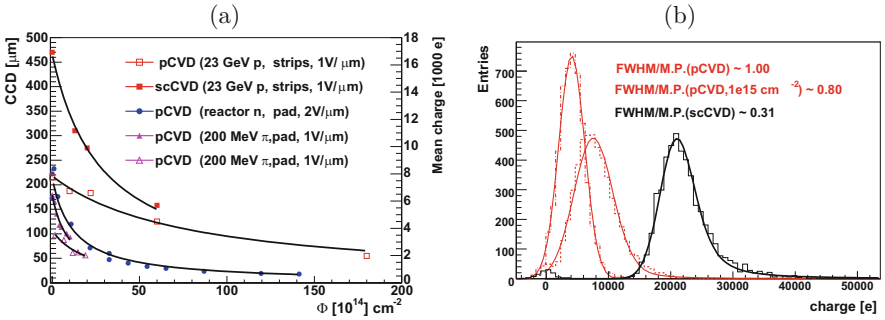


Fig. 21.37 (**a**) CCD vs. fluence of different particles. Detectors were 500 μm thick. The Eq. (21.44) is fitted to the data [89–91]. The irradiation particle, electrode geometry and electric field is given in brackets. (**b**) Energy loss distribution in CVD pad detectors. The value of FWHM, corrected for electronics noise, over most probable energy loss is shown

Table 21.7 Charge collection distance degradation parameter for different irradiation particles [89–92]

	70 MeV	800 MeV	23 GeV p	200 MeV π	Reactor neutrons
$K[10^{-18}\,\mu m^{-1}cm^{-2}]$	1.76	1.21	0.65	~3.5	~3 – 4

to mean charge of 2770 e_0. At lower fluences the first term dominates and for $CCD_0 \sim 200\ \mu m$ CCD only decreases by 15% after 10^{15} cm^{-2} of 23 GeV protons.

The homogeneity of the response over the detector surface, which is one of the drawbacks of pCVD detectors, improves with fluence for pCVD as the collection distance becomes smaller than grain size. For the same reason also the distribution of energy loss in pCVD detector, initially wider than in scCVD, becomes narrower (Fig. 21.37b). The energy loss distribution in pCVD diamond is Gaussian, due to convolution of energy loss distributions (Landau) in grains of different sizes.

One of the main advantages of the diamond is the fact that at close to room temperatures no annealing or reverse-annealing effects were observed.

The drawbacks of grains in pCVD detectors can be largely overcome by using **3D diamond detectors** [93], who share the same concept with silicon detectors (see Fig. 21.38). The vertical electrodes are produced by focused laser light which graphitizes the diamond. Whether the vertical electrode serves as cathode or anode depends on metal bias grid on the surface of the detector. Very narrow electrodes of ~2 μm diameter can be made along 500 μm thick device with low enough resistivity to allow good contacts and doesn't increase the noise. Such a good aspect ratio allows even smaller cell sizes than in silicon.

The first tests showed >75% charge collection efficiency in 500 μm pCVD diamond detector of ganged 150 × 150 μm^2 cells with bias voltages of only few tens Volts [94]. A much better homogeneity of charge collection over the surface (columns are parallel to grain boundaries) and narrower distributions of collected charge were obtained than in planar diamond detectors.

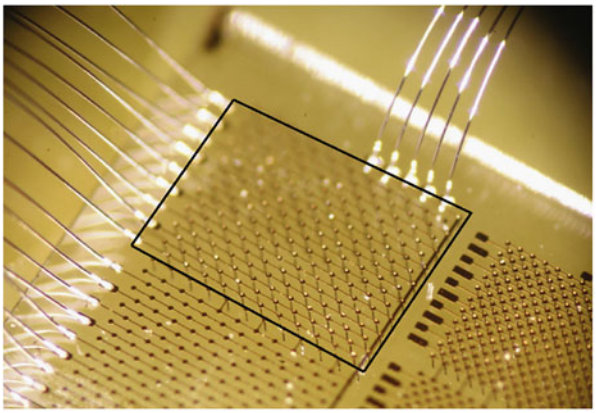


Fig. 21.38 Photograph of 3D diamond strip detector in black rectangle with a square cell of 150 μm size. Strip detector of the same geometry with planar electrodes is shown below (from [93])

The main problem with diamond 3D detectors is the rate of production in particular for large area as even if laser beam is powerful enough and is split into several parallel beams. Currently the rate is limited to roughly ten thousand holes a day.

The diamond detectors are used also outside particle physics for particle detection such as for fusion monitoring where neutrons are detected, for alpha particle detection, for determination of energy and temporal distribution of proton beams and in detection of ions during the teleradiology. They are also exploited for soft X-ray detection, where the solar-blindness and fast response of diamond detectors are the keys of their success.

21.4.4 Other Semiconductor Materials

Silicon is in many respects far superior to any other semiconductor material in terms of collected charge, homogeneity of the response and industrial availability. Other semiconductor materials can only compete in niche applications where at least one of their properties is considerably superior or where the existing silicon detectors cannot be used. For example, if low mass is needed or active cooling can not be provided, high leakage current in heavily irradiated silicon detectors is intolerable and other semiconductor detectors must be used.

The growth of compound semiconductors is prone to growth defects which are frequently unmanageable and determine the properties of detectors before and after irradiation. If a high enough resistivity can be achieved, the detector structure can be made with ohmic contacts. However, it is more often that either a Schottky contact or a rectifying junction is used to deplete the detector of free carriers. Only

a few compound semiconductors have reached a development adequate for particle detectors. They are listed in the Table 21.2. For particle physics application some other very high-Z semiconductor such as CdZnTe or HgI_2 are inappropriate due to large radiation length.

Silicon Carbide was one of the first alternatives to silicon proposed in [95, 96]. It is grown as epitaxial layer or as bulk material. Even though at present the latter exhibits a lot of dislocations (inclusions, voids and particularly micro-pipes) in the growth and the former is limited to thicknesses around 50 μm, both growth techniques are developing rapidly and wafers are available in large diameters (10 inch). Due to the properties similar to diamond the same considerations apply as for diamond with an important advantage of 1.4 larger specific ionization (55 e-h/μm).

Presently the best performing detectors are produced by using slightly n-doped epitaxial layers of ≈50 μm forming a Schottky junction.They exhibit 100% charge collection efficiency after full depletion and negligible leakage current [97]. Also detectors processed on semi-insulating bulk (resistivities $\sim 10^{11}$ Ωcm) with the ohmic contacts show CCD up to 40 μm [96] at 1 V/μm for few hundred μm thick material.

After irradiation with hadrons the charge collection deteriorates more than in silicon or in diamond. For epitaxially grown SiC the degradation of CCD is substantial with $K_e \approx 20 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ and $K_h \approx 9 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ for reactor neutron and 23 GeV p irradiated samples at high electric fields of 10 V/μm [97, 98]. The leakage current is unaffected by irradiation or it even decreases [98].

GaAs The resistivity of GaAs wafers is not high enough for the operation with ohmic contacts and detectors need to be depleted of free carriers, which is achieved by Schottky contact or a $p-n$ junction. GaAs detectors were shown to be radiation hard for γ-rays (^{60}Co) up to 1 MGy [99]. As a high Z material these detectors are very suitable for detection X and γ rays.

Their tolerance to hadron fluences is however limited by loss of charge collection efficiency, which is entirely due to trapping of holes and electrons. The V_{fd} decreases with fluence [100, 101] which is explained by removal or compensation of as grown defects by irradiation. Although larger before the irradiation, the trapping distance of electrons shows a larger decrease with fluence than the trapping distance of holes. The degradation of charge collection distance at an average field of 1 V/μm (close to saturation velocity) in 200 μm thick detectors is very large $K_{e,\pi} \approx 30 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ and $K_{h,\pi} \approx 150 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ [100]. One should however take into account that specific ionization in GaAs is four times larger than in diamond.

The leakage current increases moderately with fluence up to few $10\,\mathrm{nA/cm^2}$, much less than in silicon, and starts to saturate at fluences of around $10^{14}\,\mathrm{cm^{-2}}$ [100, 101]. The GaAs exhibit no beneficial nor reverse annealing of any detector property at near to room temperatures.

GaN The GaN detectors produced on few μm thin epitaxial layer shown charge collection degradation which is much larger than in Si [102]. Further developments in crystal growth may reveal the potential of material.

21.4.5 Comparison of Charge Collection for Different Detectors

The key parameter relevant to all the semiconductor particle detectors is the measured induced charge after passage of minimum ionizing particles. The charge collection dependence on fluence in different semiconductor pad detectors is shown in Fig. 21.39a. A 3D pad detector (all columnar electrodes connected together) shows best performance, while smallest charge is induced in SiC and pCVD diamond detectors. The induced charge decreases with fluence and at most few thousand e_0 can be expected at $\Phi_{eq} > 10^{16}\,\mathrm{cm}^{-2}$.

Although pad detectors are suitable for material comparison the effect of segmentation and choice of the type of the read-out electrodes determine to a large extent the performance of the detectors. The superior charge collection performance of segmented silicon planar detectors with n^+ electrodes to pad detectors can be seen in Fig. 21.39b. A signal of around 7000 e_0 is induced in epitaxial p-type and Fz p-type strip detectors at $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$. At the highest fluence shown the signal in a silicon pad detector is only half of that in a strip detector. On the other hand a device with p^+ readout performs worst of all.

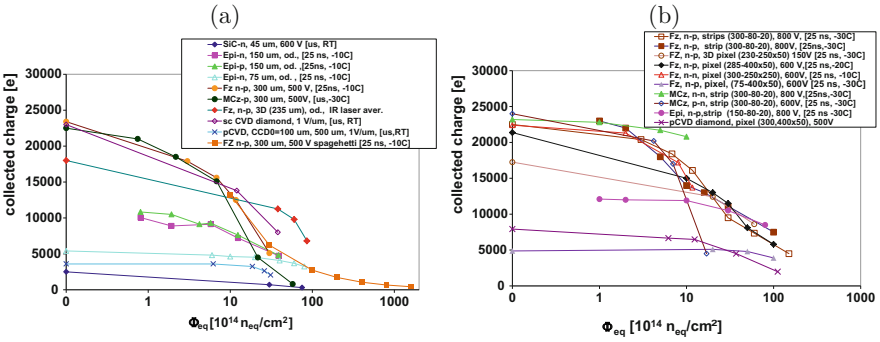


Fig. 21.39 (**a**) Comparison of charge collection in different detectors and materials; given are material, thickness, voltage, [shaping time of electronics and temperature]. "od." means at $V_{bias} > V_{fd}$. All detectors were irradiated with 23 GeV protons, except 75 μm epi-Si and 300 μm thick "spaghetti" diode which where irradiated with reactor neutrons. For diamond detectors the mean, not the most probable, charge is shown. (**b**) Charge collection in different segmented devices; the segmentation is denoted for strips (thickness-pitch-width) and pixels (thickness-cell size)]. Solid markers denote neutron irradiated and open 23 GeV proton irradiated samples

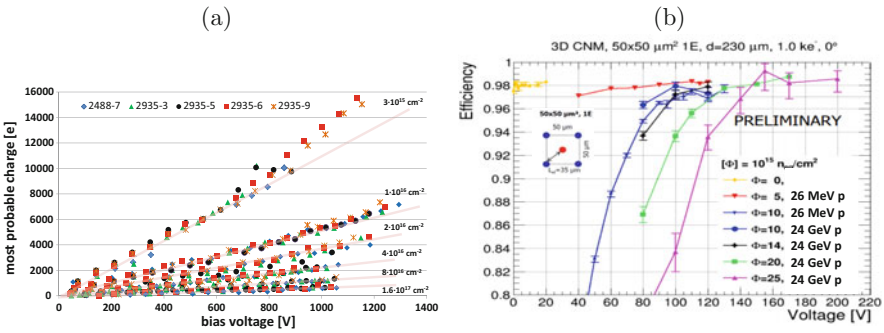


Fig. 21.40 (**a**) Dependence of collected charge in different planar silicon detectors on voltage up to the extreme fluences at −10 °C. The color bands are to guide the eye. (**b**) Test beam (120 GeV π) measurement of detection efficiency in heavily irradiated 3D detector with single electrode cell of $50 \times 50 \mu m^2$ shown in the inset [79]

The detection efficiency, however, depends on signal-to-noise ratio, which should be maximized. A choice of material, electrode geometry and thickness determine the electrode capacitance which influences the noise of the connected amplifier (Chap. 10). At given pixel/strip geometry the highest electrode capacitance has a 3D silicon detector, followed by a planar detector with n^+ electrodes, due to required p-spray or p-stop isolation which increases the inter-electrode capacitance. Even smaller is the capacitance of p^+ electrodes which is of 1 pF/cm order for strip detectors. The smallest capacitance is reached for diamond detectors owing to small dielectric constant.

21.4.5.1 Operation at Extreme Fluences

A combination of trapping times saturation, active neutral bulk and charge multiplication allows silicon detectors to be efficient in radiation environments even harsher than that of HL-LHC, approaching those of FCC. The operation of silicon detectors was tested up to $\Phi_{eq} = 1.6 \cdot 10^{17}\,\mathrm{cm}^{-2}$ and is shown in Fig. 21.40a for short strip detectors with ganged electrodes ("spaghetti" diode). Detectors remained operational and most probable charge of around 1000 e was measured in 300 µm thick detectors at 1000 V. At high fluences ($>2 \cdot 10^{15}\,\mathrm{cm}^{-2}$) the collected charge is linearly proportional with o bias voltage in whole range of applicable voltages and the dependence of charge on voltage and fluence can be parametrized with only two free parameters [103],

$$Q(V, \Phi_{eq}) = k \cdot V \cdot \left(\frac{\Phi_{eq}}{10^{15}\ \mathrm{cm}^{-2}}\right)^b, \tag{21.45}$$

where $b = -0.683$ and $k = 26.4$ e/V for 300 µm thick detectors.

A small cell size 3D detector ($50 \times 50 \mu m^2$, 1E) irradiated with charged hadrons to $\Phi_{eq} = 2.5 \cdot 10^{16}\,cm^{-2}$ was recently found to be fully efficient at voltages even below 200 V (see Fig. 21.40b) [79]. A rough simulation of collected charge in such a device based on known data predicts collected charge $>3000\ e_0$ after the fluence of $10^{17}\,cm^{-2}$, which may be already enough also for successful tracking.

21.4.6 Radiation Damage of Monolithic Pixel Detectors

The monolithic pixel detectors, which combine active element and at least first amplification stage on the same die, are widely used in x-ray and visible imaging applications. Their use as particle detectors is limited for applications where radiation environments are less severe (space applications, $e^+ - e^-$ colliders), either because of small hadron fluences or because of radiation fields dominated by leptons and photons (see Fig. 21.3). The CCD is the most mature technology while CMOS active pixel sensors were successfully used for particle detection in STAR experiment at Relativistic Heavy Ion Collider over the last decade. These detectors are more susceptible to radiation damage due to their charge collection mechanism and the readout cycle. Recently several CMOS foundries offered a possibility to apply high voltage which can be used for depletion of substrate on which CMOS circuitry resides thereby enabling fast charge collection by drift. This greatly enhanced both radiation hardness of CMOS detectors and their speed.

The principles of operation of these detectors were addressed in section on Solid state detectors. Here on only the aspects of radiation hardness of aforementioned detectors will be addressed.

21.4.6.1 CCDs

The CCD[5] is intrinsically radiation soft. The transfer of the charge through the potential wells of the parallel and serial register is very much affected by the charge loss. At each transfer the fraction of the charge is lost. The charge collection efficiency is therefore calculated as $CCE = (1 - CTI_p)^n \times (1 - CTI_s)^m$, where CTI_s and CTI_p denote the charge collection inefficiency of each transfer in serial and parallel register. An obvious way of improving the CCE is a reduction of the number of transfers (m and n). Applications requiring high speed such as ILC, where the readout of the entire detector ($n \sim 2500$) within 50 μs is neeeded, the serial register is even omitted ($m = 0$) and each column is read-out separately (column parallel CCD [104]). The CCDs suffer from both surface and bulk damage.

[5]CCD is often not considered to be monolithic devices.

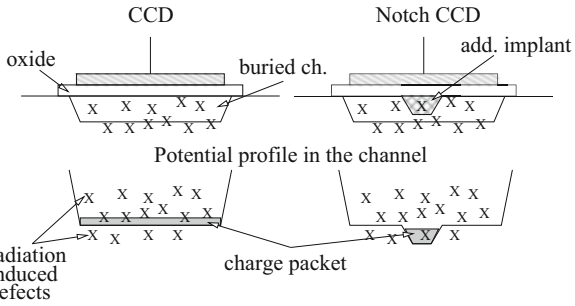


Fig. 21.41 The principle of notch CCD. An additional n^+ implant creates the minimum in potential

The increase of CTI is a consequence of bulk and interface traps. There are several methods to improve the CTI:

- The transport of the charge takes place several hundred nm away from the surface by using a n^+ implant (buried channel), which shifts the potential minimum. The transport is less affected by trapping/detrapping process than at the interface traps.
- Operation of CCDs at low temperatures leads to filling of the traps with carriers—electrons. Since emission times are long (Eq. (21.16)) the amount of active traps is reduced.
- If the density of signal electrons (n_s) is larger than the trap concentration only a limited amount of electrons can be trapped, thus $CTI \propto N_t/n_s$. An additional n^+ implant can be used for buried channel CCDs to squeeze the potential minimum to much smaller volume (see Fig. 21.41).
- The CTI depends on the charge transfer timing, i.e. on the clock shapes. The transfer of the charge from one pixel to another should be as fast as possible to reduce the trapping. The choice of the clock frequency, number of the phases (2 or 3) and shape of the pulses, which all affect the CTI, is a matter of optimization (see Fig. 21.42). The transfer time from one well to another can be enhanced by an implant profile which establishes gradient of the electric field.
- If traps are already filled upon arrival of the signal charge they are inactive and CTI decreases. The effect can be achieved either by deliberate injection of charges (dark charge) or by exploiting the leakage current. In the same way also the pixel occupancy affects the CTI.

The radiation affects also operation of detectors due to surface effects. The surface generation current which is a consequence of interface traps is in most of the applications the dominant source of current in modern CCDs. Very rarely the bulk damage is so high that the bulk generation current dominates. The surface dark current can be greatly suppressed by inverse biasing of the Si-SiO_2 interface [105].

The voltage shift due to oxide charge requires proper adjustment of the amplitude of the gate drive voltages. However the supply current and power dissipation of the gate drivers can exceed the maximum one as they both depend on the square of the voltage amplitude.

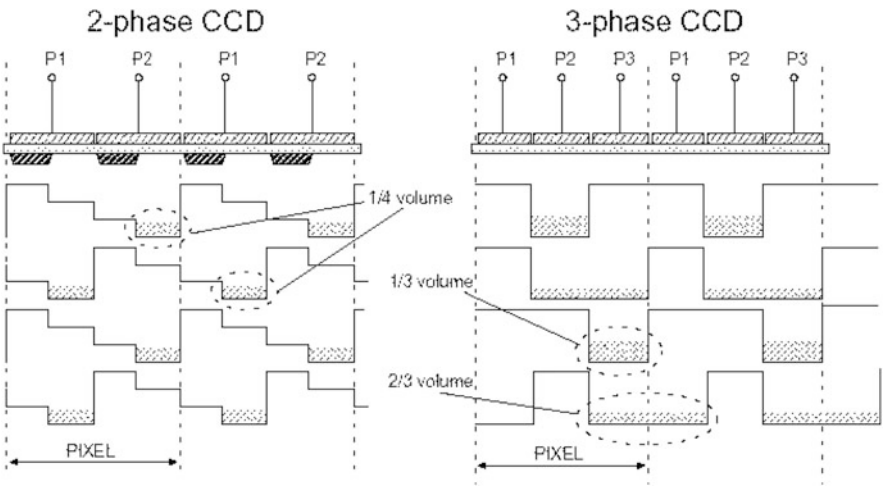


Fig. 21.42 Comparison of charge transfer of 2 and 3 phase CCD (P1,P2,P3 denote gate drive voltages). Note that the potential well occupies 1/4 of the pixel volume for 2 phase CCD and up to 2/3 for 3-phase CCD. The signal charge is shaded

The CCDs can probably not sustain radiation fields larger than at the ILC (see Table 21.1), particularly because of the bulk damage caused by neutrons and high energy leptons.

21.4.6.2 Active CMOS Pixels

In conventional monolithic active CMOS pixel sensors (Chap. 5) [106, 107] the n^+ well collects electron hole pairs generated by an ionizing particle in the p doped epitaxial layer (see Fig. 21.43). The built-in depletion around the n^+ well is formed enabling the drift of the carriers. In the major part of the detector the charge is collected from epitaxial p-type silicon through the diffusion. The charge collection process depends on epitaxial layer thickness and takes tens of ns. Above 90% of the cluster charge is induced within ~100 ns for 15 μm thick epitaxial layer [108]. Since the n^+ wells are used as collection electrodes, only nMOS transistors can be used for the signal processing circuit. The level of complexity of signal processing after the first stage depends on the CMOS technology used (number of metal layers, feature size).

The charge collection by thermal diffusion is very sensitive to electrons lifetime, which decreases due to the recombination at deep levels. The loss of collection efficiency and consequently smaller signal-to-noise ratio is the key limitation for their use. The way to increase the radiation tolerance is therefore the reduction of diffusion paths. This can be achieved by using many n^+ collection diodes per pixel area, which improves the charge collection efficiency. The price for that is a larger

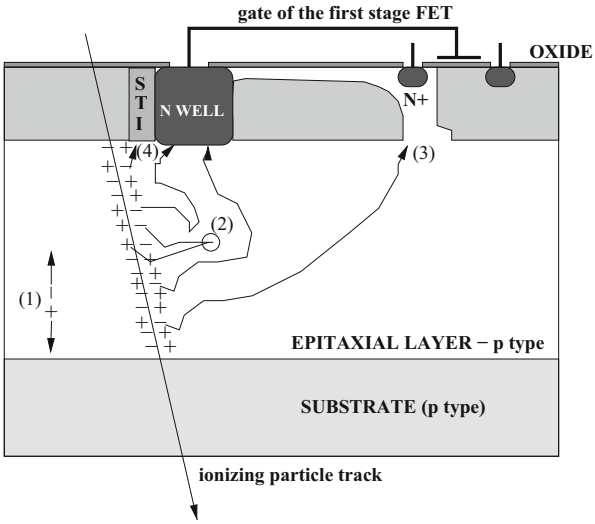


Fig. 21.43 Schematic view of the radiation effects in active CMOS Pixel Detector: (1) generation of carriers—leakage current; (2) recombination of diffusing carriers; (3) positive oxide charge buildup leads to punch trough the p well; (4) charge trapping at shallow trench isolation structure

capacitance and leakage current of the pixel. An increase of the epitaxial layer will increase the fraction of recombined charge, but the absolute collected charge will nevertheless be larger. The reduction of the collection time can be achieved by a gradual change of epitaxial layer doping concentration which establishes electric field.

The generation-recombination centers give rise to the current and cooling is needed to suppress it. It increases the noise and requires more frequent reset of the pixel.

The active pixel detectors were proven to achieve detection efficiencies of >95% at $\Phi_{eq} = 2 \cdot 10^{12}\,\mathrm{cm}^{-2}$ [109], suggesting an upper limit of radiation tolerance to hadron fluences of $\Phi_{eq} \leq \cdot 10^{13}\,\mathrm{cm}^{-2}$.

The active pixel sensors are CMOS circuits and therefore susceptible to surface damage effects. Apart from the damage to transistor circuitry which is discussed in next section in some CMOS processes the n-well is isolated from p-well by shallow trench SiO_2 isolation. The radiation induced interface states serve as trapping centers and reduce the signal. The active pixel sensors were shown to be tolerant to ionizing radiation doses of up to 10 kGy [110]. The damage effects discussed above are shown in Fig. 21.43.

Depleted CMOS

In recent years a so called depleted CMOS or high voltage CMOS (HV-CMOS) process has become available by different foundries. These processes allow application of high voltage to the p substrate which becomes depleted. Charge collection by drift significantly improves both speed and radiation tolerance of these devices.

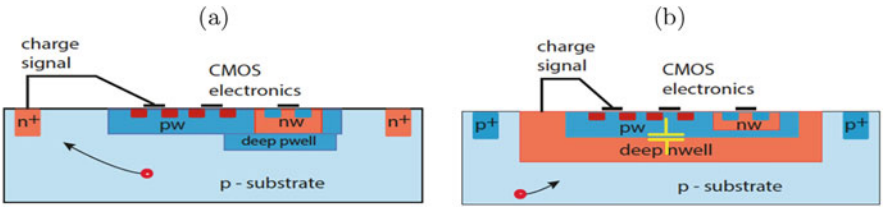


Fig. 21.44 Schematic view of (**a**) small electrode and (**b**) large electrode HV-CMOS detectors [111]

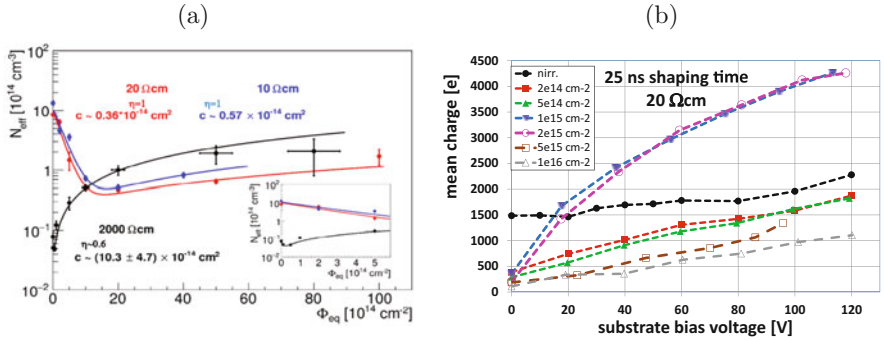


Fig. 21.45 (**a**) Dependence of substrate N_{eff} on fluence of devices produced by two different foundries on different substrate resistivities. Fit of Hamburg model to the data is shown with initial acceptor removal parameters left free [37]. (**b**) Charge collection in irradiated passive HV-CMOS diode array connected to LHC speed electronics [36]

The devices differ mostly in the way the collection electrode is realized. A small n^+ collection electrode is beneficial (see. Fig. 21.44a) for its small capacitance. If, however, a n^+ electrode is inside a large n-well the capacitance is determined by the size of n-well (see. Fig. 21.44b), but the charge collection is faster and more homogeneous. The optimum design therefore depends on the application (see Ref. [112]). Both options are under consideration for the upgrade of pixel sub-detectors at HL-LHC.

Relatively high doping concentration of the substrate (from N_{eff} =few 10^{12} to 10^{15} cm^{-3}) emphasizes the importance of effective acceptor removal with irradiation. For low resistivity substrates the removal of shallow acceptors dominates over the creation of deep ones and the effective doping concentration initially decreases with irradiation. The active/depleted thickness at given voltage increases resulting in larger collected charge. After the initial acceptors are removed the deep acceptors determine the depleted thickness regardless of the choice of initial substrate. The dependence of N_{eff} on fluence for different initial substrate resistivities/doping is shown in Fig. 21.45a [37]. The increase of active thickness is reflected also in charge collection measurements shown in Fig. 21.45b for a low resistivity, 20 Ωcm, device. Note that after the irradiations the contribution from the charges diffusing

from the undepleted substrate to the depleted region vanishes and almost no charge is measured without bias. The contribution of carriers diffusing from the undepleted substrate disappears already after $\Phi \sim 10^{14}\,\mathrm{cm}^{-2}$ [36] which is the reason for initial drop of charge collection efficiency.

Recent studies of pixelated devices established the need for metallization their backside and/or thinning them down [113]. In most processes the high voltage for depletion of the substrate is applied from the contact on top of the device (see Fig. 21.44). After irradiation the increase of resistivity of undepleted bulk can have a large impact on fraction of weighting potential traversed by the carriers and therefore induced charge. Low impedance biasing electrode (HV bias) relatively far away from the sensing electrode and long lateral drift paths of carriers in devices without back side biasing can result in smaller induced charge than expected from the active thickness.

21.5 Electronics

The front-end electronics is an essential part of any detector system. The application specific integrated circuits (ASIC) are composed of analog and digital parts. The analog part usually consists of a preamplifier and a shaping amplifier, while the digital part controls the ASIC and its communication with readout chain. The fundamental building block of the circuit, transistors, can be either bipolar or field-effect devices.

The benefits of either bipolar or field-effect transistors as the first amplifying stage are comprehensively discussed in [114] (see Chapter 6). The equivalent noise charge of the analog front end is given by

$$ENC^2 \approx 2\,e_0\,(I_{nm} + I_m M^2 F)\,\tau_{sh} + \frac{4\,k_B\,T}{g_m\,\tau_{sh}}(C_d + C_c)^2 \tag{21.46}$$

where τ_{sh} is shaping/integration time of the amplifier, I_{nm} and I_m the non-multiplied and multiplied currents flowing in the control electrode of the transistor, M current multiplication factor and F excess noise factor, C_d detector capacitance, C_c capacitance of the transistor control electrode and g_m the transconductance of the transistor. The transconductance measures the ratio of the change in the transistor output current vs. the change in the input voltage. The first term is also called current (parallel) noise while the second term is called voltage (series) noise [115]. The radiation of particle detectors therefore increases both parallel noise through I_{nm}, I_m and series noise through C_d.

The excess noise factor is determined by the gain and effective ratio of hole and electron ionization coefficients k_{eff} [116]

$$F = k_{eff}\,M + (1 - k_{eff})(2 - 1/M) \tag{21.47}$$

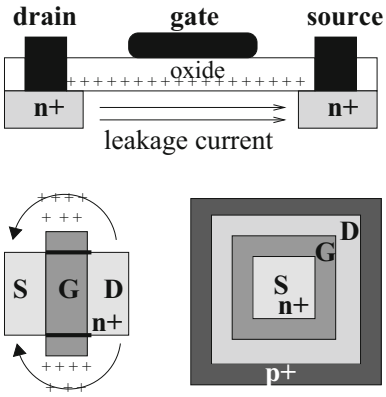


Fig. 21.46 Schematic view of the MOSFET leakage current (top). The standard FET design (bottom left) with parasitic current paths and enclosed transistor design (bottom right)

For moderate gains of $M \approx 10$ and $k_{eff} < 0.01$ usual for silicon tracking detectors $F \sim 2$. It follows from Eq. (21.46) that a voltage noise should dominate the current noise if charge multiplication should increase the signal/noise ratio. If this in not true the current noise increases faster than the signal with bias voltage. Therefore, integration time and electrode size and design should be carefully optimized.

The bipolar transistors are susceptible to both bulk and surface damage while field effect transistors suffer predominately from surface effects. It is the transconductance that is affected most by irradiation.

MOSFET

Advances in integrated electronics circuitry development lead to reduction of feature size to the deep sub-micron level in CMOS technology. The channel current in these transistors is modulated by the gate voltage. The accumulation of positive oxide charge due to ionizing radiation influences the transistor threshold voltage V_{th} (See Fig. 21.46). For nMOS transistors the channel may therefore always be open and for pMOS always closed after high doses. This is particularly problematic for the digital part of the ASIC leading to the device failure. The operation points in analog circuits can be adjusted to some extent to accommodate the voltage shifts. The threshold voltage depends on the square of the oxide thickness and with thick oxides typical for MOS technologies in the previous decades (>100 nm) the radiation hardness was limited to few 100 Gy. At oxide thickness approaching 20 nm the relation $V_{th} \propto d^2$ breaks down (see Fig. 21.8b) as explained in Sect. 21.3.2.1. The deep sub-micron CMOS processes employing such thin oxides are therefore intrinsically radiation hard. The weak dependence of V_{th} on dose for deep sub-micron CMOS processes is shown in Fig. 21.47a. The interface states introducing the leakage current are largely deactivated (see Sect. 21.3.2) in deep sub-micron CMOS transistors, leading to almost negligible surface current (Fig. 21.47b). Also mobility changes less than 10% up to 300 kGy.

Even with transistor parameters not severely affected by radiation the use of so called enclosed transistor layout (ELT) [117] is sometimes required to eliminate the

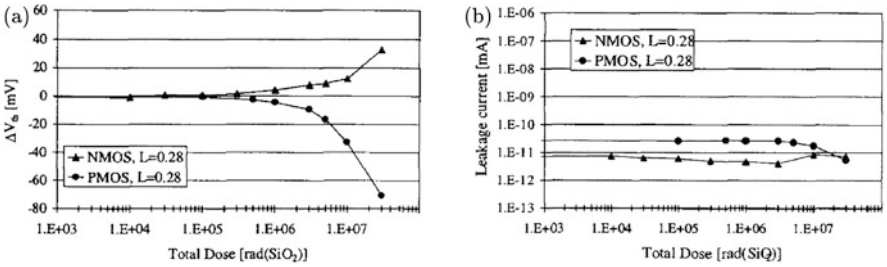


Fig. 21.47 (**a**) Threshold voltage shift of enclosed nMOS and standard pMOS transistors as a function of the total dose for a 0.25 μm technology. (**b**) Leakage current for the same transistors [117].

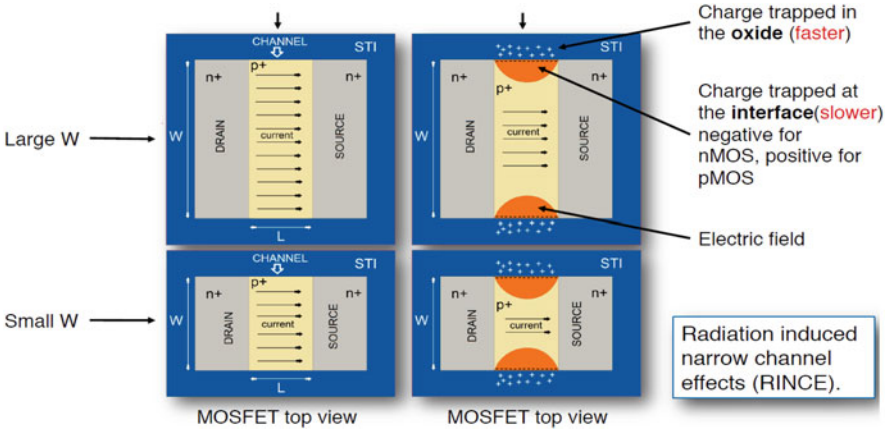


Fig. 21.48 Schematic view of Radiation-Induced Narrow Channel Effect

radiation effects on large arrays of transistors. Radiation induces transistor leakage through the formation of an inversion layer underneath the field oxide or at the edge of the active area (see Fig. 21.46). This leads to source-to-drain and inter-transistor leakage current between neighboring n^+ implants. The former can be avoided by forcing all source-to-drain currents to run under the gate oxide by using a closed gate. The inter-transistor leakage is eliminated by implementing p^+ guard rings.

Development of dedicated libraries to implement enclosed transistors for each deep-sub micron process is often too demanding or the functionality required for a given surface doesn't allow enclosed transistors. In such cases a so called Radiation-Induced Narrow Channel Effect (**RINCE**) shown in Fig. 21.48 can occur.

The positive charge trapped in the lateral shallow trench isolation oxide (STI) attracts electrons and opens a conductive channel through which leakage current can flow between source and drain. This current is usually small and [119] compared to the current that can flow in the main transistor and it only influences the subtreshold region of the transistor I–V curve. Even if the functionality of the chip is preserved

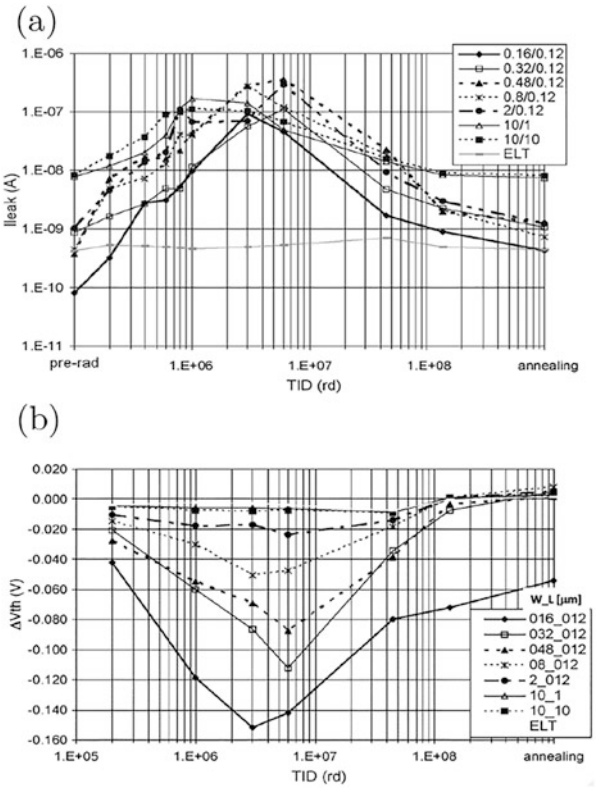


Fig. 21.49 (**a**) Evolution of the leakage current with TID for different NMOS transistor sizes (width/length in μm), up to 1.36 MGy. The last point refers to full annealing at 100 °C. The first point to the left is the pre-rad value (**b**) Same as (**a**) but showing transistor threshold voltage shift [119].

this impacts power consumption and thermal performance of the chip. At higher doses the negative charge trapped at the interface states compensates the positive space charge (NMOS transistors) and leakage current decreases (see Fig. 21.49a). Both processes of positive oxide charge and negative charge build-up at the interface states are highly dependable on dose rate, process and annealing. The increase of the transistor leakage current affected the operation of ATLAS-IBL detector [77].

Apart from parasitic leakage current the trapped oxide charges can also moderate electric field in the transistor channel particularly for narrow channel transistor where relatively larger part of the transistor is affected. If the change in threshold voltage for NMOS is small (see Fig. 21.49b), RINCE can be a bigger problem for PMOS transistors. There, positive trapped charge (holes) at the interface states adds to the positive oxide charge. As a consequence the threshold voltage and the required current to turn transistor on change significantly. As shown in Figs. 21.49 only marginal annealing effects were observed.

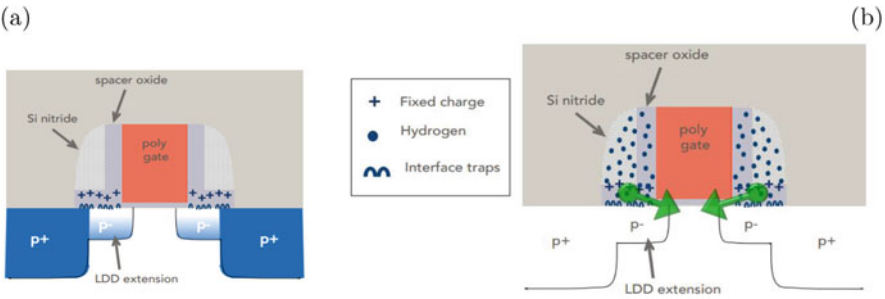


Fig. 21.50 (**a**) Radiation-Induced Short Channel Effect—charging of spacer oxide modifies free carrier concentration in LDD (p^-) layer. (**b**) Annealing releases protons/hydrogen atoms/molecules to the gate oxide [120]

Radiation Induced Short Channel Effect (RISCE) appears in both PMOS and NMOS transistors with very short channel (for both ELT and open-layout transistors) and is a consequence of transistor design with so called spacer oxide shown in Fig. 21.50a. This oxide charges and affects the amount of carriers in Low Drain Doping (LDD) extension of the transistors leading to a decrease of the transistor-on current during exposure. The radiation and temperature/annealing frees protons, neutral hydrogen atoms/molecules from spacer oxide that can reach the nearby gate oxide. There they depassivate Si-H bonds and by that change the threshold voltage and transistor-on current. This process is strongly dependent on (annealing) temperature. It can be avoided at low temperature operation ($T < 0°C$) and by switching off biasing at high temperatures [120].

The constant reduction of feature size in modern CMOS processes going from 0.35, 0.25, 0.13, 0.065, 0.045 µm have also other beneficial consequences. Ever shorter transistor channel lengths result in higher speed of the devices which consumes also less power particularly in the digital part. Larger transistor densities allow more complex signal processing while retaining the die size. Unfortunately at given power constraints, the basic noise parameters of bipolar and field-effect front-end transistors will not improve with the reduction of feature size [115].

Bipolar Transistor The main origin of damage in bipolar transistors is the reduction of minority carrier life time in the base, due to recombination processes at radiation-induced deep levels. The transistor amplification factor $\beta = I_c/I_b$ (common emitter) decreases according to $1/\beta = 1/\beta_0 + k\Phi$. The pre-irradiation value is denoted by β_0 and damage constant dependent on particle and energy by k. Since $g_m \propto \beta$, degradation of β leads to larger noise and smaller gain of the transistor. Thinner base regions are less susceptible to radiation damage, so faster transistors tend to be better.

The choice of base dopant plays an important role. A boron doped base of a silicon transistor is not appropriate for large thermal neutron radiation fields due to the large cross-section for neutron capture (3840 barns). The kinetic energy released (2.3 MeV) to Li atoms and α particles is sufficient to cause large bulk damage [118].

Single Event Effects (SEE) Unlike the bulk and surface damage, the single event effects are not cumulative. They are caused by the ionization produced by particles hitting certain elements in the circuit. According to the effect they have on operation they can be:

- **transient**; Spurious signals propagating in the circuit, due to electrostatic discharge.
- **static**; The state of memory or register bits is changed (Single Event Upset). In case of altering the bits controlling the circuit, they can disturb functionality or prevent circuits from operating (Single-event Functional Interrupt).
- **permanent**; The can destroy the circuit permanently (Single Event Latchup).

The SEE become a bigger problem with reduction of the feature size, as relatively smaller amount of ionization is required to change properties. The radiation hardening involves the use of static-RAM instead of dynamic-RAM and processing of electronics on SOI instead on silicon bulk (physical hardening). The logical hardening incorporates redundant logical elements and voting logic. With this technique, a single latch does not effect a change in bit state; rather, several identical latches are queried, and the state will only change if the majority of latches are in agreement. Thus, a single latch error will not change the bit.

21.6 Conclusions

The radiation damage of crystal lattice and the surface structure of the solid state particle detectors significantly impacts their performance. The atoms knocked-off from their lattice site by the impinging radiation and vacancies remaining in the lattice interact with themselves or impurity atoms in the crystal forming defects which give rise to the energy levels in the semiconductor band-gap. The energy levels affect the operation of any detector in mainly three ways. Charge levels alter the space charge and the electric field, the levels act as generation-recombination and trapping centers leading to increase of leakage current and trapping probability for the drifting charge. The magnitude of these effects, which all affect the signal-to-noise ratio, depends on the semiconductor material used as well as on the operation conditions.

Although the silicon, by far most widely used semiconductor detector material, is affected by all three, silicon detectors still exhibit charge collection properties superior to other semiconductors. Other semiconductors (e.g. SiC, GaN, GaAa, a-Si) can compete in applications requiring certain material properties (e.g. cross-section for incoming radiation, capacitance) and/or the crucial properties are less affected by the radiation (e.g. leakage current and associated power dissipation). Radiation effects in silicon detectors were thoroughly studied and allow for reliable prediction of the detector performance over the time in different irradiation fields.

The state of the art silicon strip and pixel detectors used at experiments at LHC retain close to 100% detection efficiency for minimum ionizing particles at hadron fluences in excess of $10^{15}\,\mathrm{cm}^{-2}$ and ionization doses of 1 MGy. The foreseen upgrade of Large Hadron Collider require hardness to even an order of magnitude larger fluences, which presently set the ultimate benchmark for operation of semiconductor particle detectors. The efforts for improving the silicon detection properties in order to meet these demanding requirements include defect engineering by adding impurity atoms, mainly oxygen, to the crystal, operation at cryogenic temperatures and placement electrodes perpendicularly to the detector surface—3D detectors.

The increase of effective doping concentration with fluence together with high voltage operation lead to charge multiplication in heavily irradiated silicon detectors which in combination with electric field in the neutral bulk and saturation of effective trapping probabilities result in efficient operation in radiation environments even harsher than that at the HL-LHC.

In recent years new detector technologies appeared, such as Low Gain Avalanche Detectors which offer along with position also time resolution and depleted CMOS monolithic detectors. The latter offer for the first time fully monolithic devices with fast response and sufficient radiation hardness. The initial dopant removal plays a crucial role in performance of both LGADs and depleted CMOS detectors. Among other semiconductors diamond is the most viable substitute for silicon in harsh radiation fields, particularly with the advent of 3D diamond detectors.

The silicon detector employed in less severe environments e.g. monolithic active pixels, charge coupled devices, silicon drift detectors are optimized for the required position and/or energy resolution and the radiation effects can be well pronounced and even become the limiting factor already at much lower doses. Longer drift and/or charge integration times increase the significance of leakage current, charge trapping and carrier recombination.

The silicon-silicon oxide border and the oxide covering the surface of silicon detectors and electronics is susceptible to ionizing radiation. The positive charge accumulates in the oxide and the concentration of interface states, acting as trapping and generation-recombination centers, increases. These effects can be effectively reduced in silicon detectors by proper processing techniques. Thin oxides ($<$20 nm) allow tunneling of electrons from the gate electrode through the oxide. They can recombine with positive charges in the oxide and also passivate interface traps. Deep sub-micron CMOS processes which utilize oxides of such thicknesses are therefore intrinsically radiation hard especially if proper design rules are used. In very deep-sub micron processes where often the use of special design rules is not possible two effects Radiation-Induced Narrow Channel Effect and Radiation Induced Short Channel Effect appear which require special adjustments in operation scenarios.

References

1. K. McKay, Phys. Rev. 76 (1949) 1537.
2. J. Kemmer, Nucl. Instr. Meth. A 169 (1980) 499.
3. M. Swartz, M. Thurston, J. Appl. Phys., 37(2) (1966) 745.
4. A.G. Holmes-Siedle and L. Adams, Radiat. Phys. Chem. 28(2) (1986) 235.
5. ATLAS Inner Detector Technical design report, CERN/LHCC/97-16, ISBN 92-9083-102-2.
6. G.D. Badhwar, Rad. Res. 148 (1997) 3.
7. W.N. Spjeldvik and P.L. Rothwell, The Earth's Radiation Belts. In: Environmental Research Paper No. 584, Air Force Geophysics Laboratory, U.S. Department of the Air Force, AFGL-TR-83-0240, Massachusetts (1983).
8. V.A.J. van Lint, T.M. Flanagan, R.E.Leadon, J.A. Naber, V.C.Rogers, *Mechanism of Radiation Effects in Electronic Materials*, John Wiley & Sons, 1980.
9. T.F. Luera et al., IEEE Trans. NS 34(6) (1987) 1557.
10. M. Huhtinen, Nucl. Instr. and Meth. A 491 (2002) 194.
11. G. Lindström, Radiation Damage in Silicon Detectors, Nucl. Instr. and Meth. A 512, 30 (2003).
12. W. de Boer, Phys. Stat. Sol. (a) 204, No. 9 (2007) 3004.
13. E. Gaubas et al., Mat. Sc. Sem. Proc. 75 (2018) 157165.
14. M. Mikuž et al., "Extreme Radiation Tolerant Sensor Technologies" presented at 26th Vertex conference, Las Caldas, Spain, September, 2017..
15. R. Radu et al., Journal Of Applied Physics Vol. 117 (16) (2015) 164503.
16. M.M. Atalla, E. Tannenbaum and E.J.Scheibner, Bell. Syst. Techn. J. 38 (1959) 749.
17. D.M. Fleetwood, J. Appl. Phys. 73 (10) (1993) 5058.
18. C.T. Sah, IEEE Trans. NS 23 (6) (1976).
19. A. Goetzberger, V. Heine, E.H. Nicollian, Appl. Phys. Lett. 12 (1968) 95.
20. A.G. Revesz, IEEE Trans. Electron Dev. ED-12 (1965) 97.
21. R. Wunstorf et al., Nucl. Instr. and Meth. A 377 (1996) 290.
22. N.S. Saks, M. G. Ancona and J. A. Modolo, IEEE Trans. NS 31(6) (1984) 1249.
23. J Zhang et al., JINST 7 (2012) C12012.
24. R.H. Richter et al., Nucl. Instr. Meth. A 377 (1996) 412.
25. W. Füsel et al., Nucl. Instr. and Meth. A 377 (1996) 177.
26. S. Ramo, Proc. I.R.E. 27 (1939) 584.
27. G. Kramberger et al., IEEE Trans. NS 49(4) (2002) 1717.
28. T.J. Brodbeck et al., Nucl. Instr. and Meth. A 395 (1997) 29.
29. G. Casse et al., Nucl. Instr. and Meth. A 487 (2002) 465.
30. R. Wunstorf, Ph.D. thesis, Hamburg University 1992, DESY FH1K-92-01 (October 1992).
31. Michael Moll, Ph.D. thesis, Hamburg University 1999, DESY-THESIS-1999-040, ISSN-1435-8085.
32. J. Adey, PhD Thesis, Univeristy of Exceter, 2004.
33. J. Adey et al., Physica B 340342 (2003) 505508.
34. G. Lindström et al.(RD48), Nucl. Instr. and Meth. A 466 (2001) 308.
35. A. Khana et al., Solar Energy Materials & Solar Cells 75 (2003) 271.
36. A. Affolder et al., JINST Vol. 11 (2016) P04007.
37. I. Mandić et al., JINST 12 P02021 2017.
38. E. Cavallaro et al., JINST 12 C01074 2017.
39. B. Hiti et al., JINST 12 P10020 2017.
40. P. Dias de Almeida et al., "Measurement of the acceptor removal rate in silicon pad diodes", 30th CERN-RD50 Workshop, Krakow, 2017.
41. G. Kramberger et al., JINST Vol. 10 (2015) P07006.
42. E. Buchanan for LHCb Velo collaboration, "The LHCb VELO & ST Operational Performance Run II", PoS (Vertex 2017) 016.

43. M. Kocian for ATLAS collaboration, "Operational Experience of ATLAS SCT and Pixel Detector", PoS(Vertex 2017) 017.
44. C. Barth for CMS collaboration, "CMS pixel and strip rad damage measurements", 31st CERN-RD50 Workshop, Geneve, 2017.
45. J. Härkönen, Nucl. Instr. and Meth. A 518 (2004) 346.
46. I. Pintilie, et al., Meth. Instr. and Meth. A 514 (2003) 18.
47. G. Kramberger et al., Nucl. Instr. and Meth. A 515, 665 (2003).
48. G.Lindström et al., Nucl. Instr. and Meth. A 568 (2006) 66.
49. G. Kramberger et al., Nucl. Instr. and Meth. A 609 (2009) 142.
50. V. Eremin et al., Nucl. Instr. and Meth. A 360 (1995) 458.
51. V. Eremin et al., Nucl. Instr. and Meth. A 476 (2002) 556.
52. G. Kramberger et al., Nucl. Instr. and Meth. A 497 (2003) 440.
53. G. Kramberger et al.,"Investigation of Irradiated Silicon Detectors by Edge-TCT ", IEEE Trans. Nucl. Sci. Vol. 57(4), 2010, p. 2294.
54. R. van Overstraetem and H.de Man, Solid-State Electronics 13(1970),583–608.
55. W. Maes, K. de Meyer, R. van Overstraetem, Solid-State Electronics 33(1990),705–718.
56. J. Lange et al.,"Charge Multiplication Properties in Highly-Irradiated Epitaxial Silicon Detectors", PoS(Vertex 2010) 025.
57. M. Koehler, IEEE Trans. NS 58 (2011) 3370.
58. I. Mandić et al. Nucl. Instr. and Meth. A603 (2009) 263.
59. G. Casse et al. Nucl. Instr. and Meth. A624 (2010) 401.
60. I. Mandić et al., JINST 8 (2013) P04016.
61. R. Mori et al., Nucl. Instr. and Meth. A796 (2015) 131.
62. I. Mandić et al.,Nucl. Instr. and Meth. 629 (2011) 101.
63. G. Kramberger et al., Nucl. Instr. and Meth. A 481 (2002) 297.
64. O. Krasel et al., IEEE Trans. NS 51(1) (2004) 3055.
65. RD50 Status Report 2008, CERN-LHCC-2010-012 and LHCC-SR-003 (2010).
66. W. Adam et al., JINST Vol. 11 (2016) P04023.
67. M. Mikuž et al.,"Extreme Radiation Tolerant Sensor Technologies", presented at 26th Vertex Conference, Las Caldas, 2017.
68. M. Moll, et al., Nucl. Instr. and Meth. A 426 (1999) 87.
69. V. Palmieri et al., Nucl. Instr. and Meth. A 413 (1998) 475.
70. K. Borer et al., Nucl. Instr. and Meth. A 440 (2000) 5.
71. V. Eremin et al., Nucl. Instr. and Meth. A 583 (2007) 91.
72. E. Verbitskaya et al., IEEE Trans. NS 49(1) (2002) 258.
73. A. Chilingarov et al., Nucl. Instr. and Meth. A 399 (1997) 35.
74. S.I. Parker et al., Nucl. Instr. and Meth. A395 (1997) 328.
75. A. Zoboli et al., IEEE Trans. NS 55(5) (2008) 2775.
76. G. Pellegrini et al, Nucl. Instr. and Meth. A 592 (2008) 38.
77. M. Backhaus et al., Nucl. Instr. and Meth. A831 (2016) 65.
78. G. Darbo et al., JINST 10 (2015) C05001.
79. J Lange et al., JINST Vol. 13 (2018) P09009.
80. M. Garcia-Sciveres et al. Nucl. Instr. and Meth. A 636 (2011) 155.
81. H. Sadrozinski et al., Nucl. Instr. and Meth. A831 (2016) 18.
82. G. Pellegrini et al., Nucl. Inst. Meth. A765 (2014) 14.
83. N. Cartigli, H. Sadrozinski and A. Seiden, REPORTS ON PROGRESS IN PHYSICS Vol. 81 (2018) 026101.
84. N. Cartiglia et al., Nucl. Instr. and Meth. A850 (2017) 83.
85. G. Kramberger et al., Nucl. Instr. and Meth. A891 (2018) 68.
86. H. Sadrozinski et al., "Properties of HPK UFSD after neutron irradiation up to 6e15 n/cm^2 " to appear in Nucl. Instr. and Meth. A (2018).
87. M. Ferrero et al., Nucl. Instr. and Meth. A919 (2019) 16.
88. G. Kramberger et al., Nucl. Instr. and Meth. A898 (2018) 53.
89. W. Adam et al., Nucl. Instr. and Meth. A447 (2000) 244.

90. H. Kagan et al., Nucl. Instr. and Meth. A582 (2007) 824.
91. M. Mikuž et al., "Study of polycrystalline and single crystal diamond detectors irradiated with pions and neutrons up to 3 × 1015 cm-2", IEEE Nucl. Sci. Symp. Conference Record, San Juan (2007) N44-5.
92. N. Venturi et al., "Results on Radiation Tolerance of Diamond Detectors", presented at 11th International Hiroshima Symposium on the Development and Application of Semiconductor Tracking Detectors, Okinawa, Japan, (2017), to appear in Nucl. Instr. and Meth. A.
93. F.Bachmair et al., Nucl. Instr. and Meth. A786 (2015) 97.
94. N. Venturi et al., JINST 11 (2016) C12062.
95. F. Nava et al., Nucl. Instr. and Meth. A437 (1999) 354.
96. M. Rogala et al., Nucl. Phys. B 78 (1999) 516.
97. S. Sciortino et al., Nucl. Instr. and Meth. A552 (2005) 138.
98. F. Moscatelli et al., IEEE Trans. NS 53(4) (2006) 1557.
99. S.P. Beaumont et al., Nucl. Instr. and Meth. A322 (1992) 472.
100. R.L. Bates et al., Nucl. Instr. and Meth. A395 (1997) 54.
101. M. Rogala et al., Nucl. Instr. and Meth. A410 (1998) 41.
102. J. Grant et al., Nucl. Instr. and Meth. A576 (2007) 60.
103. G. Kramberger et al., JINST 8 (2013) P08004.
104. C.J.S. Damerell et al., Nucl. Instr. and Meth. A512 (2003) 289.
105. N.S. Saks et al., IEEE Trans. NS 27 (1980) 1727.
106. G. Claus et al., Nucl. Instr. and Meth. A465 (2001) 120.
107. R. Turcheta et al., Nucl. Instr. and Meth. A458 (2001) 677.
108. G. Deptuch et al., Nucl. Instr. and Meth. A465 (2001) 92.
109. M. Deveaux et al., Nucl. Instr. and Meth. A583 (2007) 134.
110. M. Deveaux et al., Nucl. Instr. and Meth. A552 (2005) 118.
111. N. Wermes and H. Kolanoski, *Teilchendetektoren: Grundlagen und Anwendungen*, Springer Spektrum (2016), ISBN 978-3-662-45349-0.
112. H. Pernegger et al., JINST 12 (2017) P06008.
113. I. Mandić et al., Nucl. Instr. and Meth. A903 (2018) 126.
114. V. Radeka et al., Ann. Rev. Nucl. Part. Sci. 37 (1988) 217.
115. H. Spieler, *Semiconductor Detector Systems*, Oxford University Press, Oxford (2005) ISBN 0-19-852784-5 TK9180. S68 2005.
116. R. J. MaIntyre, IEEE Trans. Elec. Devices Vol. 13 (1966) 164.
117. P. Jarron et al., Nucl. Phys. B 78 (1999) 625.
118. I. Mandić et al., Nucl. Instr. and Meth. A518 (2004) 474.
119. F. Faccio and G. Cervelli, IEEE Trans. Nucl. Sci. Vol. 57 (2005) 2413.
120. F. Faccio et al., IEEE Trans. Nucl. Sci. Vol. 65 (2018) 164.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Chapter 22
Future Developments of Detectors

Ties Behnke, Karsten Buesser, and Andreas Mussgiller

22.1 Introduction

Large scale detectors in particle physics take many years to plan and to build. The last generation of large particle physics detectors for the energy frontier, ATLAS and CMS, have been operating for more than 10 years, and upgrades for them are now being done. Studies for the next generation of experimental facilities have been ongoing for a number of years. In this section future directions in integrated detector design are discussed, as they were visible at the time of writing this report.

At the moment the biggest approved project in particle physics is the upgrade of the Large Hadron Collider (LHC) towards high luminosity running. This project is scheduled to be completed by 2027, and major upgrades to the two main collider detectors ATLAS and CMS are planned. Beyond the LHC, an electron-positron collider has been discussed for many years, to fully explore the Higgs and the top sector and to complement the discovery reach of a hadron machine at the energy frontier with a high precision program.

The requirements as far as detectors are concerned are very different for these two types of projects: for the LHC luminosity upgrade fundamental changes to the underlying philosophy of the existing detectors are not possible, but significant technological development is needed to meet the challenges of extreme radiation environments and high event rates. For a yet not existing electron-positron collider a detector can be designed from ground up, optimised to meet the ambitious physics agenda of such a facility.

T. Behnke (✉) · K. Buesser · A. Mussgiller
DESY, Hamburg, Germany
e-mail: ties.behnke@desy.de

© The Author(s) 2020

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_22

Several strategy discussions at national and international levels have consistently put a high energy electron positron collider far up on the list of future projects in the field [1–4]. Such a facility should serve as a Higgs factory, run at least at an energy of 250 GeV, but should also provide an upgrade path towards the top threshold and beyond. With the results from the current run of the LHC which show no indications of direct signs for new physics, the role of ultimate precision especially at the Higgs production threshold has been much strengthened [5].

The International Linear Collider, ILC, is a mature project to build an electron-positron collider, which could eventually push into the TeV regime, realised as a linear accelerator. The facility is described in the Technical Design Report from 2012 [6], and targets an initial energy of 250 GeV, upgradable to 1 TeV. To reach energies in the multi-TeV range in an electron-positron collider, another technology will be needed. The CLIC technology, developed mostly at CERN, is a promising candidate for such a machine [7, 8].

With the strong emphasis on precision Higgs physics, circular machines have become again a subject of study. A circular collider like the FCC-ee project, pursued at CERN [9], could reach the Higgs and possibly the top pair threshold in a ring of around 100 km circumference. Such a facility could also be used for the next large hadron collider, reaching energies of up to 100 TeV [10]. A similar project, CEPC/SppC, is under discussion in China [11–13].

A number of smaller projects are currently pursued in the field of experimental high-energy physics as well, for example, the B-factory at KEK, or long baseline neutrino experiments like Dune.

22.2 Challenges at Future Colliding Beam Facilities

The Large Hadron Collider, LHC, has seen first beams in 2008. Until 2018, a spectacular physics harvest has taken place, with the undisputed highlight the discovery of the long-sought-after Higgs particle in 2012. The energy of the collider has reached its design value, and the collider will continue to run in this configuration for another approximately 5 years, until 2024.

Already now the LHC has exceeded its design luminosity of $10^{34}\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and is expected to accumulate a total integrated luminosity of $300\,\mathrm{fb}^{-1}$ in the first running phase ("Phase-I") that extends to 2024. This will result in significant new insights into the physics of the electroweak symmetry breaking, and significant new information on physics beyond the standard model.

During the Phase-II, starting around 2027 and extending to 2035 or beyond, the LHC will increase its luminosity by about a factor of 10. ATLAS and CMS will extend their physics reach [15] significantly with this upgrade. The discovery reach for supersymmetric particles for example will be extended by some 20–30%, access to rare decay modes e.g. of the Higgs Boson will be improved, and flavour changing neutral currents through top decays might become accessible. Many other measurements will profit from this improvement as well. However, the increased

luminosity is payed for with more severe background conditions, with a much larger number of events per beam crossing, and a resulting challenge to the sub-detectors. In particular the innermost detectors will need major upgrades, together with the readout and data acquisition systems, to handle the new conditions.

It should be noted here that also studies have been initiated for detectors of possible future very large hadron colliders that could succeed the LHC and explore energy ranges of up to 100 TeV. One such concept of a hekto-TeV hadron collider is discussed within the framework of the FCC-hh study at CERN [10], another, SPPC, is part of the CEPC study in China [12]. The requirements for the detectors of such machines are just being explored and are far from being fully understood. The main challenges are related to the large jet energies and boosted event topologies, that require very large magnetic fields, large detector dimensions, and highly segmented detectors. In addition, the radiation environment is harsh and requires very radiation hard detectors.

An electron-positron collider like the ILC or FCC-ee poses different but unique challenges to its detectors. It puts a premium on precision physics, particularly on the precision reconstruction of jet masses. The experimental environment is benign by LHC standards, which allows one to consider technologies and solutions which have not been possible during the development of the LHC detectors.

To reach high precision in the overall reconstruction of event properties, each sub-system must reach excellent precision by itself. In addition, however, in the combination of sub-systems into a complete detector extreme care has to be taken to be able to fully utilise the precision of the sub-detectors. Among the most relevant parameters is the amount of dead material, in particular for the inner tracking detectors, and its radiation hardness. Low mass detectors are a key requirement, and add a major challenge to the system. High readout-speed is another ingredient, without which the high luminosity of the collider can not be exploited fully.

An experiment at an electron-positron collider has to be designed to extract maximum information from the event, and utilise the available luminosity as much as possible. It has to be able to reconstruct as many different topologies and final states as possible. This implies that the focus of the development has to be the reconstruction of hadronic final states, which are by far the most numerous ones in nearly all reactions of interest. A typical event topology is a multi-jet final state, with typical jet energies of order of 50–100 GeV. In contrast to the LHC, where many collisions occur in one bunch crossing, typically only one event of interest takes place at the linear collider, even at very high luminosities. With well below 100 particles per jet the total number of particles in the final state is comparatively small. This makes it possible to attempt the reconstruction of every single particle, neutral or charged, in the event. A major focus of the detector development therefore will be the capability of the detector to identify individual particles as efficiently as possible, and to reconstruct the properties of each particle as precisely as possible. This has large implications for the overall design of the detector.

Even though the event topology at an electron-positron collider is intrinsically clean, and there are no underlying events nor multiple interactions, as they are present in a hadron collider, backgrounds nevertheless do play a role. In particular

for the innermost and the most forward systems, beam induced backgrounds are significant. Electron-positron pairs created in the interaction of the two highly charged bunches add significant background to the event, and detectors close to the beam need to be able to cope with these. This background is particularly relevant at linear colliders, which, due to the smaller repetition rate of the interactions, need to focus their beams very strongly at the interaction region to reach the luminosity goals. Circular electron-positron colliders on the other hand can operate with less strongly focussed beams, since they re-use the beams after each turn, operating at much larger repetition rates.

22.3 Hadron Colliders

The LHC and its envisaged upgrade to the HL-LHC provides a physics program well until the middle of the 2030s. As discussed above, plans for the next colliders at the energy frontier are being made already now. A possible far-future option is a very large hadron collider. Recently, the conceptual design report for the Future Circular Collider (FCC), a ≈100 km long storage ring proposed for CERN, has been published. The proposal foresees to start with an e^+e^- collider for Higgs precision studies (FCC-ee [9]) that could be replaced by a hadron collider, the FCC-hh [10], at a later stage, probably not before the 2060s. Table 22.1 summarises the basic parameters of HL-LHC and FCC-hh in comparison to the LHC.

The LHC detectors are operating since quite some time now and are very well understood. This experience helped to design the upgrades that are required to cope with the challenges of the oncoming LHC luminosity upgrade, as will be discussed in the next Sect. 22.3.1. The FCC-hh challenges to the detectors are quite different; first concepts for detectors are under discussion and will be presented in Sect. 22.3.2.

Table 22.1 Some basic design parameters of the LHC, HL-LHC and FCC-hh (nominal) [10]

Parameter	Unit	LHC	HL-LHC	FCC-hh
Center-of-mass energy	TeV	14	14	100
Peak luminosity/IP	10^{34} cm^{-2} s^{-1}	1	5	30
Number of bunches	#	2808	2808	10,400
Bunch population	10^{11}	1.15	2.2	1.0
Time between bunches	ns	25	25	25
Beam spot size at IP	µm	16.7	7.1	3.5
Bunch length	cm	7.55	7.55	8
Accelerator length	km	27	27	97.75
Peak pile-up events/bunch crossing	#	25	130	950
Pile-up line density	mm^{-1}	0.2	1.0	8.1
Pile-up time density	ps^{-1}	0.1	0.29	2.43
Total ionising dose at $r = 2.5$ cm	MGy	1.3	13	270

22.3.1 Detector Upgrades for the High-Luminosity-LHC

The two major colliding beam experiments at the LHC, ATLAS and CMS, have recorded large data sets starting in 2010. The currently installed innermost detectors were designed to cope with track densities and to withstand the radiation doses expected during the LHC Phase-I running that extends until 2024. For the high-luminosity operation phase of the LHC both experiments will replace their inner tracking detector with completely new systems.

The tracking detectors of both large LHC experiments are mostly based on silicon technology detectors. Over the past years, an intense R&D effort has taken place, to re-design and re-optimise the inner detectors for both ATLAS and CMS. Fundamentally no changes in technology will take place, both detectors will rely on an all-silicon solution for the tracking. In addition, ATLAS will remove the transition radiation detector from its system, and extend its silicon tracker to larger radii. Owing to the track trigger concept, CMS completely re-designs its tracker and utilises novel detector modules that allow for an on-module p_T discrimination of charged particle tracks. Both Phase-II trackers will again follow a classical barrel and end cap design. However, compared to the Phase-I trackers, ATLAS will use wedge-shaped sensor modules in its end caps of the tracker, whereas CMS will rely on rectangular modules in this part of the detector. Both future trackers will have substantially increased granularity to cope with the expected pile up of up to 200 events per bunch crossing, and very much improved radiation tolerance, which will significantly go beyond the one of the Phase-I detectors and suffice for operation throughout the Phase-II era.

The amount of insensitive material is a significant performance limiting factor of the current trackers, both at ATLAS and at CMS. The large amount of material in the present trackers not only reduces the performance of the trackers themselves, but also has a negative impact on the performance of the electromagnetic calorimeters directly outside of the tracking systems. The reduction of material is therefore another important goal of the tracker upgrades. CMS will use 320 μm thick sensors with an active thickness of 200 μm, as compared to 500 μm thickness in the present detector, novel structural materials, and novel powering and cooling schemes will make this goal achievable.

For the innermost layers of the future trackers radiation tolerance will be of even larger importance than today. Current technologies are not able to withstand the anticipated rates for longer periods. A number of novel technologies are under consideration, 3D Silicon pixel sensors or diamond tracking detectors. Even solutions which do not involve Silicon—like Micromegas trackers—are being discussed.

The higher rates at the upgraded LHC will not only challenge the hardware of the tracker, but also put large demands on the readout and the trigger system. In particular, the latter will have to be significantly upgraded to handle the anticipated rates without a loss of sensitivity. The tracker might well play a central role here, as the early trigger on track-like objects already in the level-1 trigger will significantly

reduce the trigger rate. Triggering on tracks rather than just increasing the trigger thresholds will maintain a much better sensitivity to a broad range of signals, in particular for the much sought-after new physics signals.

The final layout is based on the concept of a "long pixel" detector. In this Ansatz the pixel size is increased compared to current pixels to something like $100\,\mu\text{m} \times 2\,\text{mm}$. It appears possible to keep the power per pixel constant compared to current pixel readouts, thus resulting in a tracker which has a channel count larger by two orders of magnitude than the current strip trackers, but a similar overall power consumption.

Although the tracking detectors are most affected by the increased luminosity, other detectors will be affected as well. The calorimeters will see much increased backgrounds in the forward direction, which might necessitate upgrades or significant changes. A serious problem might be that the ATLAS liquid argon calorimeter in the forward direction heats up under the backgrounds to a point where is will no longer function. In this case—which will only be known once operational experience under real LHC conditions is available—the replacement by a warm forward calorimeter might be necessary. CMS intends to make major changes to its calorimeter system, replacing the hadronic section and part of the electromagnetic section with a highly granular calorimeter, using technology which has been developed and will be described later in the section on detectors at electron positron colliders. For all detectors the capability to handle larger rates will be needed, and might make updates and replacements of the readout electronics necessary. This even applies to parts of the muon system, again primarily in the forward direction. ATLAS e.g. is considering to replace the drift tubes in the forward direction with ones of smaller diameter, to limit the occupancy. In any case upgrades to the trigger and the data acquisition are needed.

22.3.1.1 Novel Powering Schemes

The minimisation of power consumption will play a central role in the upgrades of the LHC tracker detectors for the LHC Phase-II. Traditionally readout electronics are the main generators of heat in the detectors, which needs to be cooled away. Both ATLAS and CMS employ sophisticated liquid cooling systems, operating at pressures below the atmospheric pressure, to cool away approximately 33 kW from the tracking detector alone. Power is brought to the electronics at low voltages, typical for semiconductor operation. The resulting large cross sections of conductors add significantly to the overall material of the detector.

Several alternative schemes are under consideration, to limit the material and volume needed by the power lines. In one approach, called DC-DC, a large voltage is provided at the frontend. For the same power delivered a significant reduction of the amount of copper needed can be obtained. On the front-end the larger voltage is then transformed to the needed lower voltage. An optimised method to transform

the voltages without large power loss, and without large and bulky circuitry, is the subject of intense R&D.

An alternative option is serial powering. Here as well power is supplied to the front-end at a high potential. By putting several readout circuits in series, the power is reduced at each chip to the needed level. This approach promises reduced power loss and less material at the detector, but presents the experimenter with problems of proper grounding of the detector elements. By putting systems in series potentially a correlation between chips due to changing power consumption levels may be introduced. This method as well is the subject of intense R&D.

22.3.1.2 Novel Mechanical Structures and Cooling

The all-silicon trackers developed for the ATLAS and CMS operation at LHC Phase-II conditions rely on sophisticated mechanical systems, which are light-weight and at the same time provide the necessary precision and services to the detector modules. They need to be able to operate at low temperatures, and withstand thermal cycles with a temperature differential of up to 50°.

In contrast to previous designs, where cooling and positioning of modules was achieved via separate features of the mechanical structures, the new designs will combine these functionalities in single features with the goal of substantially reducing the amount of passive materials in the tracker volume. In addition, bi-phase evaporative CO2 cooling will be used as coolant, which not only has a larger radiation length X_0 compared to conventional coolants, but also allows to use pipes with smaller diameters and wall thickness, which even further reduces the material budget. However, smaller pipe diameters require significant improvements in the type of heat spreaders that are used to transport the heat from the source to the coolant. Due to their thermal properties carbon foams are widely used for this purpose. They provide a relatively large thermal conductivity at low mass. Moreover, carbon foams can be tuned to the specific needs of an application, by adjusting the pore-size and the amount of carbon deposited on the cell structure, which defines both the density of the foam and its thermal conductivity. Figure 22.1 shows a microscopic image of a stainless steel cooling pipe embedded in a block of carbon foam. In the sample shown the heat transfer between foam and cooling pipe is established via a layer of Boron Nitride doped glue that is pushed into the open-pore foam.

Support structures for silicon tracking devices are typically made of carbon fibre reinforced polymer (CFRP), which—due the demand of high stiffness rather than high strength—employ high or even ultra-high modulus carbon fibres. These fibres have the positive side effect that carbon fibres with a high Young modulus typically also have a large thermal conductivity in fibre direction, which is beneficial for cooling the detector or is even actively used for cooling. As the HL-LHC trackers are designed for an integrated luminosity of up to $4000\,\mathrm{fb}^{-1}$ over a operation time of 12 years without maintenance and several thermo cycles, longevity and in particular moisture uptake is a concern for the mechanical support systems.

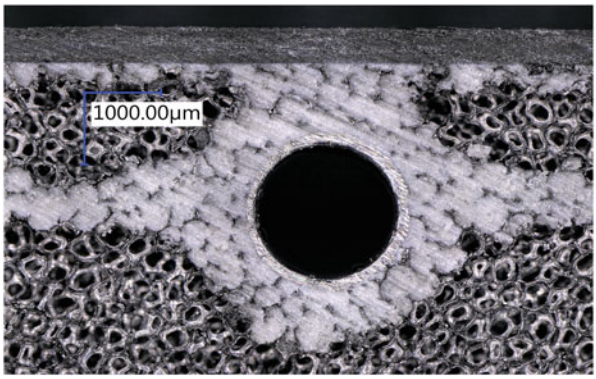


Fig. 22.1 Stainless steel cooling pipe embedded in a block of Carbon foam (credit DESY)

CFRPs with cyanate ester based resin systems are known for their low moisture uptake, however, recent industrial developments show that epoxy based systems have similar behaviour with the advantage of longer shelf life times and thus easier use of the raw material.

In general machining of CFRP with the precision required for e.g. positioning of the sensitive detector modules is not feasible, especially for layouts with small number of layers. The designs of tracker support structure therefore often follow the paradigm of "precision by glueing". The positioning elements requiring high precision machining and placement are made from e.g. Aluminium or PEEK plastic and placed on a jig prior to the assembly. The CFRP parts are then glued to these positioning elements resulting in a stiff and precise support structure. With this design and production method the tolerances on the machining and production of the used CFRP can be relaxed, which eases the production process and reduces cost while maintaining the quality of the final support structure.

22.3.2 Emerging Detector Concepts for the FCC-hh

FCC-hh will pose new challenges to the detectors [10]. A 100 TeV proton collider has not only discovery potential, given by the increased energy compared to LHC, but will also provide precision measurements as the cross sections for Standard Model (SM) processes in combination with the high luminosity lead to large event samples [13]. The envisaged detector concepts must therefore be able to measure multi-TeV jets, leptons, and photons from heavy resonances as well as Standard Model processes with high precision. As the established SM particles are small in mass, compared to the 100 TeV CMS energy of the collider, event topologies will be heavily boosted into the forward directions. A further challenge are the expected simultaneous pp collisions in one bunch crossing ('pile-up') that are expected to

reach numbers of 1000 at the FCC-hh, significantly above what is seen at LHC (60) and expected for HL-LHC (200). In particular, the anticipated separation between vertices of pile-up events is of the same order as the multiple scattering effect on the tracker vertex resolution, which renders resolving pile-up with classical 3D tracking nearly impossible. A promising approach to overcome this problem is to use 4D tracking by adding precise timing information to the tracker hits and exploiting the time structure of the pile-up events. For its HL-LHC operation the CMS experiment is foreseeing this approach already by introducing of the so-called MIP Timing Detector (MTD) that will be installed directly after the future tracker and provide timing information with a resolution of about 30 ps [14].

A reference detector for FCC-hh has been defined that at this time does not represent a specific choice for the final implementation, but rather serves as a concept for the study of physics potential and subsystem studies [10]. Figure 22.2 shows a rendering of the reference detector together with a quadrant view that shows the coverages in $|\eta|$. The detector has an overall length of 50 m and a diameter of 20 m. The central detector covers the regions of $|\eta| \leq 2.5$. Two forward

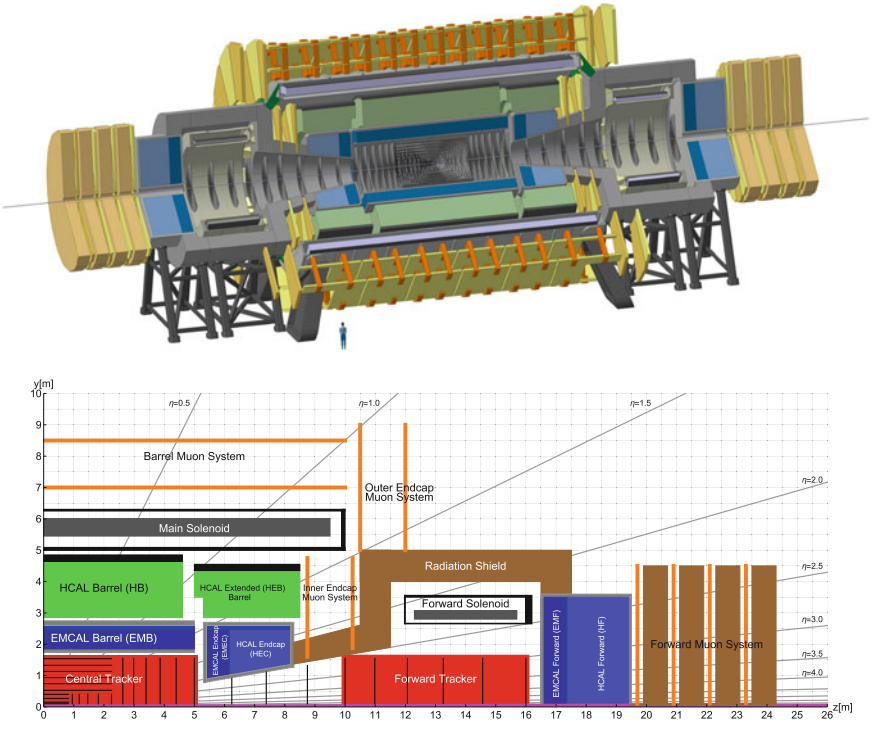


Fig. 22.2 The FCC-hh reference detector (top) has an overall length of 50 m and a diameter of 20 m. The quadrant view (bottom) shows the main detector elements and the coverage in $|\eta|$. Both figures from [10] (credit CERN/CC BY 4.0)

spectrometers cover rapidity regions of up to $|\eta| \approx 4$. A central detector solenoid with an inner bore of 10 m delivers a field of 4 T for the central regions. Two options are under study for the forward spectrometer magnets, either solenoids or dipoles. No iron return yokes are foresee, as the necessary amount of iron would be very heavy and expensive. As a consequence, the magnetic stray fields in the detector cavern will be significant which raises the need for separate service caverns some distance away.

The central tracker extends to a radius of 1.6 m. The calorimeter system consists of a LAr electromagnetic calorimeter with a thickness of 30 radiation lengths and a scintillator-iron based hadronic calorimeter of 10.5 nuclear interaction lengths. A muon system is foreseen for the outer and forward parts of the detector.

A significant challenge for the FCC-hh detector will be the radiation levels. Figure 22.3 (top) shows the expected total ionising dose rate in the detector components after a total luminosity of 30 ab^{-1} has been integrated. It is expected that the total rate for the inner tracking layers would accumulate to about 300 MGy. The radiation levels in the hadronic calorimeters would be at about 6–8 kGy,

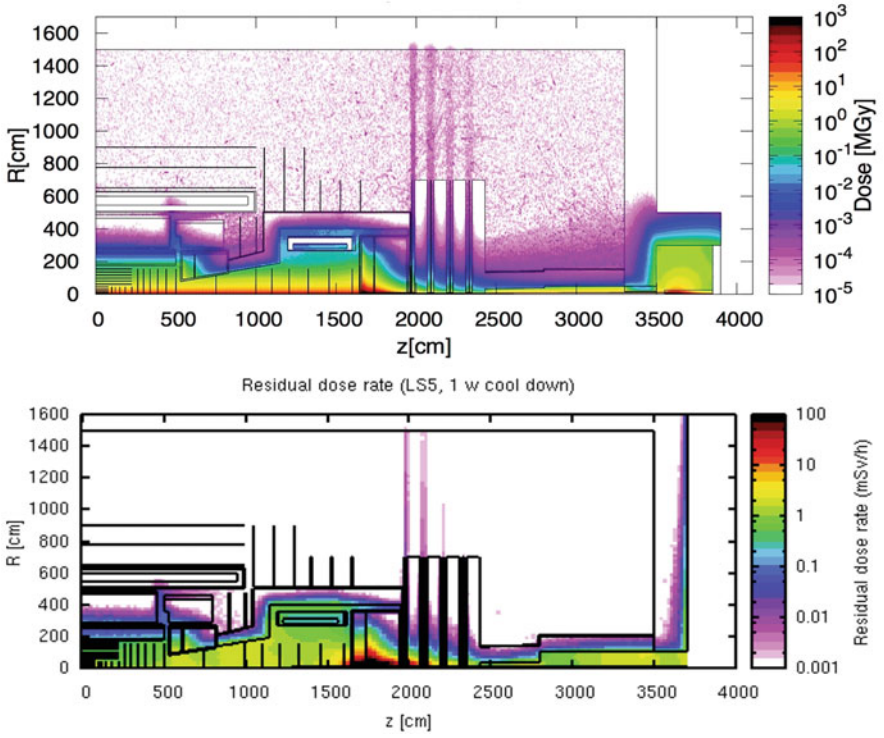


Fig. 22.3 Top: Total ionising dose for 30 ab^{-1} of integrated luminosity. Bottom: Radiation dose after one week of cool-down towards the end of the FCC-hh operation [10] (credit CERN/CC BY 4.0)

which is below the limiting number for the use of organic scintillators. Figure 22.3 (bottom) shows the radiation dose rate after one week of cool-down time towards the end of FCC-hh operations. The resulting dose rates of about 1 mSv/h in the tracker volume put limitations on person access for maintenance purposes.

22.4 Electron-Positron Colliders

The realisation of high energy electron-positron collisions has been the subject of many studies over the last years. Two fundamentally different options exist: a large circular collider, as e.g. proposed in form of the FCC-ee at CERN, or a linear collider. Due to synchroton radiation losses a circular collider is limited in its energy reach. The FCC proposal, with a ring of about 100 km in circumference, could reach with acceptable losses a final energy enough to reach the top-pair production threshold. It is economically not very sensible to go beyond this energy stage.

A linear accelerator on the other hand is intrinsically capable to reach higher energies, by extending the length of the accelerator. Over the last 20 years several technologies have been developed which promise to reach a centre-of-mass energy of 1 TeV. The international linear collider, ILC, uses superconducting cavities, a by now well established and mature technology. A fully costed design has been published in 2012 [16]. With the successful completion of the construction of the European XFEL, a large system based on the same technology has been build and successfully commissioned, providing a solid basis for estimating both costs and risks associated with this technology. An artist's drawing of the ILC facility is shown in Fig. 22.4.

To reach even higher energies the superconducting technology is not very well suited, as the achievable accelerating gradients are limited and, thus, the systems

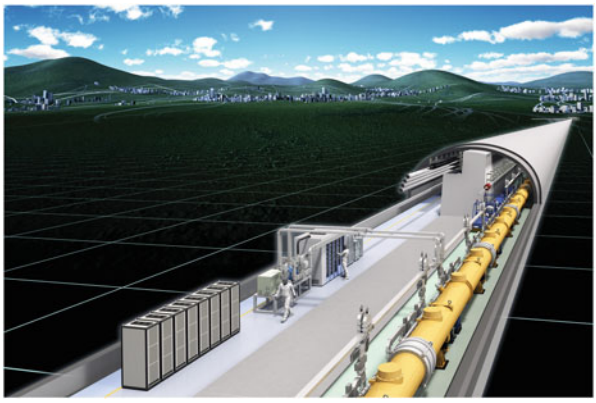


Fig. 22.4 Artist's view of the ILC tunnel in Japan. Credit: Rey Hori/KEK

will become too large. An option based on normal conducting cavities and an innovative two-beam acceleration scheme is under development at CERN in the context of the CLIC collaboration. Even more ambitious projects like plasma accelerators are being discussed as well, but are far from being available for large scale systems [17].

Politically, Japan has been discussing to come forward and host the ILC. At the time of writing this report, no final decision has been reached.

At the core of the ILC are superconducting radio-frequency cavities, made from Niobium, which accelerate the beams. After many years of intense research and development the Tesla Technology Collaboration (TTC, [18]) has developed these cavities and industrialised their production. About 800 such cavities are used in the European Free Electron Laser, built at DESY, the European XFEL [19]. Here an average acceleration gradient of 23.5 MV/m has been reached routinely, with most cavities exceeding the design value by far, almost reaching ILC design requirements. For the ILC a gradient of 31.5 MV/m is anticipated, which, however, at the time of writing this report seems to be in reach, but has not yet been realised for large numbers of cavities nor in an industrial type series production environment. Recently, an intense R&D efforts has been started to further increase the reachable gradients in superconducting RF structures. Nitrogen doping, discovered at Fermilab [20], is one subject of study, as are alternative shapes of the cavities, optimisation of the preparation of the Niobium material, and other ideas. It is hoped that the results from this R&D, which is however not the subject of this review, will significantly reduce the cost of the ILC project.

The ILC facility poses many additional challenges to the accelerator builder, which are being attacked in an intense and long-term research and development (R&D) program. The preparation of low emittance beams, the production of high intensity polarised positron beams, and the final focus of the high energy beams down to nanometer spot sizes are just some of these [21].

The key parameters of the proposed ILC facility are summarised in Table 22.2. With the current knowledge from the LHC, the importance of a high luminosity run at the Higgs threshold is strongly stressed, which led to the re-definition of the first stage of the ILC as a 250 GeV collider [24]. This also results in a significant cost saving for this first stage, an important consideration for the political discussions taking place in Japan and elsewhere. Such a collider could be realised in a tunnel infrastructure of about 20 km length. In Japan a promising site in the north of the country has been identified, which is under close scrutiny at the moment. However, it should be noted that no official decision has been reached by Japan neither on hosting the ILC, nor on its location within Japan.

The CLIC accelerator is based on normal conducting cavities, operated at 12 GHz, reaching gradients of between 80 and 120 MV/m. It is based on a novel 2-beam acceleration scheme, where one high power, low energy beam is used to produce the radio frequency needed to accelerate the high energy beam. The feasibility of this technology is investigated at CERN at the CLIC Test Facility. Over

Table 22.2 Some basic design parameters of the ILC (250 and 500 GeV options [22, 24]), CLIC (3 TeV option) [8] and FCC-ee (240 GeV parameters) [9]

Parameter	Unit	ILC-250	ILC-500	CLIC	FCC-ee
Center-of-mass energy	GeV	250	500	3000	240
Peak luminosity/IP	$cm^{-2}s^{-1}$	1.35×10^{34}	1.79×10^{34}	5.9×10^{34}	8.5×10^{34}
Pulse rate	Hz	5	5	50	
Pulse length	μs	727	727	0.24	
Number of bunches/pulse	#	1312	1312	312	328/beam
Time between bunches	ns	554	554	0.5	994
Beam size (horizontal) at IP	nm	516	474	40	13,748
Beam size (vertical) at IP	nm	7.7	5.9	1	36
Bunch length at IP	μm	300	300	44	3150
Electron polarisation	%	>80	>80	>80	0
Positron polarisation (optional)	%	>30	>30	>30	0
Accelerator length	km	20.5	33.4	≈50	97.756
Total site AC power	MW	129	164	589	308

the last year significant progress was made on demonstrating the CLIC technology (see [8] and references therein). However a major limitation remains the lack of a significant demonstration setup, which would allow full system tests in a sizeable installation.

In recent years efforts to study a circular collider option have intensified. Both at CERN and in China designs are being developed for a circular collider, based in a tunnel of about 100km in circumference, which could host an electron positron collider. The technology for such a collider is available and does not provide unsurmountable challenges, apart from the scale of the project. A design study is currently ongoing, led by CERN, to develop a conceptual design report for a such a collider hosted in the Geneva area [9]. A similar study, CEPC, is led by IHEP in Beijing, about hosting this collider in China [11]. A circular collider at the Higgs threshold would be able to deliver integrated luminosities which are—at the Higgs production threshold—higher by a factor >5 for the same running time and one interaction region, as a linear collider. It could also serve more than one interaction region simultaneously in the recirculating beams, adding up the integrated luminosities of each installed experiment. This is a big advantage over a linear collider, where the colliding beams are used only once and disposed off in beam dumps after the collision. On the other hand, the infrastructure for a 100 km installation becomes very challenging, and the energy reach of a circular machine is limited due to the losses by synchrotron radiation. It is clear that any electron-positron collider that goes beyond 350 GeV has to be linear. In that respect, linear colliders do scale with energy while circular colliders do not.

For the experimenter however the challenges at any of the proposed electron-positron collider facilities are similar. The biggest difference between the proposals is the distance between bunches. At the ILC and FCC-ee (at the Higgs threshold) this time difference is with a few 100 ns very benign. At CLIC bunch distances at sub-ns

level are anticipated, and pose additional challenges to the experiment. Nevertheless, the goals for all facilities are the same: the experiment should be able to do precision physics, even for hadronic final states, should allow the precise reconstruction of charged and neutral particles, of secondary vertices. It has to function with the very large luminosity proposed for these machines, including significant backgrounds from beam-beam interactions.

22.4.1 Physics at an LC in a Nutshell

The design of a detector at a large facility like the ILC or CLIC can not be described nor understood without some comprehension about the type of measurements which will be done at this facility. A comprehensive review of the proposed physics program at the ILC facility can be found in [23, 25, 26], a review of CLIC physics is available under [7, 8].

The discussion in this section concentrates on the physics which can be done at a facility with an energy below 1 TeV. In recent years, the physics reach of a facility operating at around 250 GeV has been closely scrutinised, both at ILC and at CLIC (which is proposed to run in an initial energy stage at 380 GeV). Earlier studies have looked at the science case for a 500 GeV machine, and have explored the additional measurements possible if an energy upgrade up to 1 TeV might be possible.

At a center-of-mass energy of 250 GeV the ILC will be able to create Higgs bosons in large numbers, mostly in the so called Higgs-Strahlungs process. Here a Higgs boson is produced associated to a Z boson. The great power of this process is that by reconstructing the Z, and knowing the initial beam energies, one can reconstruct the properties of the Higgs boson without ever looking at the Higgs boson itself. Thus a model independent and decay mode-blind study of the Higgs particle will become possible. In addition through the reconstruction of exclusive final states for the Higgs particle, high precision measurements of the branching ratios will be possible. On its own, precisions on the most relevant branching ratios of around 1% will be possible. Combined with the results from the LHC, this precision can be pushed to well below the percent level. Samples of the heavy electroweak bosons, W and Z, a focus of the program at the LEP collider, will in addition be present in large numbers, and might still present some surprises if studied in detail.

If the energy of the facility can be increased to above 350 GeV, top-quark pairs can be produced thus turning the ILC into a top factory. Again, due to the cleanliness of the initial and the final states, high precision reconstruction of the top and its parameters will become possible. A precision scan of the top pair production threshold would determine the top mass with a statistical error of 27 MeV [27], which relates to a relative precision of $\approx$0.015%, far better than what can be done at LHC.

Operating at 500 GeV or slightly above, the ILC will gain access to the measurement of the top-Higgs coupling, and start to become sensitive to a measurement

of the Higgs self coupling. This latter experiment might provide evidence for the existence of this interaction at 500 GeV, but would vastly profit from even higher energies. At 1 TeV the Higgs self coupling could be measured to within 10%, which allows for reconstructing the Higgs potential and, thus, testing a cornerstone of the predictions of the standard model and the Higgs sector. Together these measurements will allow a complete test of the Higgs sector, and thus a in-depth probe of the standard model in this unexplored region.

There are good reasons to assume that the standard model is only an effective low-energy theory of a more complex and rich theory. A very popular extension of the standard model is supersymmetry, which predicts many new states of matter. Even though the LHC sofar has not found any evidence for supersymmetry, many models exists which predict new physics in a regime mostly invisible to the LHC. Together ILC and LHC would explore essentially the complete phase space in the kinematic regime accessible at the energy of the ILC.

Should a new state of matter be found at either the LHC or the ILC, electron-positron collisions would allow to study this sign of new physics with great precision.

In addition to direct signs of new physics, as represented by new particles, the ILC would allow to indirectly explore the physics at the Terascale through precision measurements, up to energy scales which in many cases are equivalent if not higher than those at the LHC. It might well be, if no new physics is found at the LHC, that these precision measurements at comparatively low energies are our only way to learn more about the high-energy behaviour of the standard model, and to point at the right energy regime where new physics will manifest itself.

Even though the ILC has been at the focus of the discussions in this chapter, all other electron-positron collider options will have a very similar physics reach—for those energies which are reachable at each facility.

22.5 Experiments at a Lepton Collider

As discussed in the previous section, high energy lepton collisions offer access to a broad range of scientific questions. A hallmark of this type of colliding beam experiments is the high precision accessible for many measurements. A detector at such a facility therefore has to be a multi-purpose detector, which is capable to look at many different final states, at many different signatures and topologies. In this respect the requirements are similar to the ones for a detector at a hadron collider. The direction in which an lepton collider detector is optimised however is very different. Lepton collider detectors are precision detectors—something which is possible because the lepton collider events are comparatively clean, backgrounds are low, and rates are small compared to the LHC. The collision energy at the lepton collider is precisely known for every event, making it possible to measure missing mass signatures with excellent precision. This will make it possible to measure masses of supersymmetric particles with precision, or, in fact, masses of

any new particle within reach of the collider. The final states are clean and nearly background-free, making it possible to determining absolute branching ratios of essentially every state visible at the lepton collider. The reconstruction also of hadronic final states is possible with high precision, opening a whole range of states and decay modes which are invisible at a hadron machine due to overwhelming backgrounds.

This results in a unique list of requirements, and in particular on very high demands on the interplay between different detector components. Only the optimal combination of different parts of the detector can eventually deliver the required performance.

Many of the interesting physics processes at an LC appear in multi-jet final states, often accompanied by charged leptons or missing energy. The reconstruction of the invariant mass of two or more jets will provide an essential tool for identifying and distinguishing W's, Z's, H's, and top, and discovering new states or decay modes. To quantify these requirements the di-jet mass is ofter used. Many decay chains of new states pass through W or Z bosons, which then decay predominantly into two jets. To be able to fully reconstruct these decay chains, the di-jet mass resolution should be comparable or better than the natural decay width of the parent particles, that is, around 2 GeV for the W or Z:

$$\frac{\Delta E_{di-jet}}{E_{di-jet}} = \frac{\sigma_m}{M} = \frac{\alpha}{\sqrt{E\,(\mathrm{GeV})}}, \tag{22.1}$$

where E denotes the energy of the di-jet system. With typical di-jet energies of 200 GeV at a collision energy of 500 GeV, $\alpha = 0.3$ is a typical goal. Compared to the best existing detectors this implies an improved performance of around a factor of two. It appears possible to reach such a resolution by optimally combining the information from a high resolution, high efficiency tracking system with the ones from an excellent calorimeter. This so called particle flow ansatz [28, 29] is driving a large part of the requirements of the LC detectors.

Table 22.3 summarises several selected benchmark physics processes and fundamental measurements that make particular demands on one subsystem or another, and set the requirements for detector performance.

22.5.1 Particle Flow as a Way to Reconstruct Events at a Lepton Collider

Particle flow is the name for a procedure to optimally combine information from the tracking system and the calorimeter system of a detector, i.e. to fully reconstruct events. Particle flow has been one of the driving forces in the optimisation of the detectors at a Lepton Collider.

Typical events at the LC are hadronic final states with Z and W particles in the decay chain. In the resulting hadronic jets, typically around 60% of all stable

Table 22.3 Sub-Detector performance needed for key LC physics measurements (from [30])

Physics process	Measured quantity	Critical system	Critical detector characteristic	Required performance
$ZHH\ HZ \to q\bar{q}b\bar{b}$ $ZH \to ZWW^*$ $\nu\bar{\nu}W^+W^-$	Triple Higgs coupling Higgs mass B($H \to WW^*$) $\sigma(e^+e^- \to \nu\bar{\nu}W^+W^-)$	Tracker and calorimeter	Jet energy resolution, $\Delta E/E$	$3 \sim 4$ % or $30\%/\sqrt{E}$
$ZH \to \ell^+\ell^- X$ $\mu^+\mu^-(\gamma)$ $ZH + H\nu\nu \to \mu^+\mu^- X$	Higgs recoil mass luminosity weighted E_{cm} B($H \to \mu^+\mu^-$)	Tracker	Charged particle momentum resolution, $\Delta p_t/p_t^2$	5×10^{-5}
$HZ, H \to b\bar{b}, c\bar{c}, gg$ $b\bar{b}$	Higgs branching fractions b quark charge asymmetry	Vertex detector	Impact parameter, δ_b	$5\,\mu\text{m} \oplus$ $10\,\mu\text{m}/p(\text{GeV/c})\sin^{3/2}\theta$

particles are charged, slightly less than 30% are photons, only around 10% are neutral long lived hadrons, and less than 2% are neutrinos. At these energies charged particles are best re-constructed in the tracking system. Momentum resolutions which are reached in detectors are $\delta p/p^2 \approx 5 \times 10^{-5}\,\mathrm{GeV}^{-1}$, much better than any calorimeter system at these energies. Electromagnetic energy resolutions are around $\delta E_{em}/E = 0.15/\sqrt{E}(\mathrm{GeV})$, typical resolutions achieved with a good hadronic calorimeter are around $\delta E_{had}/E = 0.45/\sqrt{E}(\mathrm{GeV})$. Combining these with the proper relative weights, the ultimate energy resolution achievable by this algorithm is given by

$$\sigma^2(E_{jet}) = w_{tr}\sigma_{tr}^2 + w_\gamma \sigma_\gamma^2 + w_{h^0}\sigma_{h^0}^2, \tag{22.2}$$

where w_i are the relative weights of charged particles, photons, and neutral hadrons, and σ_i the corresponding resolution. Using the above mentioned numbers an optimal jet mass resolution of $\delta E/E = 0.16/\sqrt{E}(GeV)$ can be reached. This error is dominated by the contribution from the energy resolution of neutral hadrons, assumed to be $0.45/\sqrt{E}(GeV)$. This formula assumes that all different types of particles in the event can be individually measured in the detector. This implies that excellent spatial resolution in addition to the energy resolution is needed. Thus fine-grained sampling calorimeters are the only option currently available which can deliver both spatial and energy resolution at the same time. This assumption is reflected in the resolution numbers used above, which are quoted for modern sampling type calorimeters. Even though an absorption-type calorimeter—for example a crystal calorimeter as used in the CMS experiment—can deliver better energy resolution, it falls significantly behind in the spatial resolution, thus introducing a large confusion term in the above equation.

Formula 22.2 describes a perfect detector, with perfect efficiency, no acceptance holes, and perfect reconstruction in particular of neutral and charged particles in the calorimeter. In reality a number of effects result in a significant deterioration of the achievable resolution. If effects like a final acceptance of the detector, missing energy e.g. from neutrinos etc. is included, this number easily increases to $25\%/\sqrt{E}$ [31]. All this assumes that no errors are made in the assignment of energy to photons and neutral hadrons. The optimisation of the detector and the calorimeter in particular has to be done in a way that these wrong associations are minimised.

From the discussion above it is clear that three effects are of extreme importance for a detector based on particle flow: as good hadronic energy resolution as possible, excellent separation of close-by neutral and charged particles, and excellent hermeticity. It should also be clear that the ability to separate close-by showers is more important than ultimate energy resolution: it is for this reason that total absorption calorimeters, as used e.g. in the CMS experiment, are not well suited for the particle flow approach, as they do not lend themselves to high segmentation.

Existing particle flow algorithms start with the reconstruction of charged tracks in the tracking system. Found tracks are extrapolated into the calorimeter, and linked with energy deposits in there. If possible, a unique assignment is made between a track and an energy deposit in the calorimeter. Hits in the calorimeter belonging

to this energy deposit are identified, and are removed from further considerations. The only place where the calorimeter information is used in the charged particle identification is in determining the type of particle: calorimeter information can help to distinguish electrons and muons from hadrons. A major problem for particle flow algorithms are unassigned clusters, and mis-assignments between neutral and charged deposits in the calorimeter. The currently most advanced particle flow algorithm, PandoraPFA, tries to minimise these effects by a complex iterative procedure, which optimises the assignments, goes through several clean-up steps, and tries to also take the shower sub-structure into account [31].

What is left in the calorimeter after this procedure is assumed to have come from neutral particles. Clusters in the calorimeter are searched for and reconstructed. With a sufficiently high segmentation both transversely and longitudinally, the calorimeter will be able to separate photons from neutral hadrons by analysing the shower shape in three dimensions. A significant part of the reconstruction will be then the reconstruction of the neutral hadrons, which leave rather broad and poorly defined clusters in the hadronic calorimeter system.

Particle flow relies on a few assumptions about the event reconstruction. For it to work it is important that the event is reconstructed on the basis of individual particles. It is very important that all charged tracks are found in the tracker, and that the merging between energy deposits in the calorimeter and tracks in the tracker is working as efficiently as possible. Errors in this will quickly produce errors for the total energy, and in particular for the fluctuations of the total energy measured. Not assigning all hits in the calorimeter to a track will also result in the creating of additional neutral clusters, the so called double counting of energy. Reconstructing all particles implies that the number of cracks and the holes in the acceptance should be minimised. This is of particular importance in the very forward direction, where the reconstruction of event properties is complicated by the presence of backgrounds. However, small errors in this region will quickly introduce large errors in the total energy of the event, since many processes are highly peaked in the forward direction.

In Fig. 22.5 the performance of one particular particle flow algorithm, PandoraPFA [31] is shown, as a function of the dip angle of the jet direction, $\cos\theta$. The performance for low energies of the jets, 45 GeV is close to the optimally possible resolution if the finite acceptance of the detector is taken into account. At higher energies particles start to overlap, and the reconstruction starts to pick up errors in the assignment between tracks and clusters. This effect, called confusion, will deteriorate the resolution, and will increase at higher energies. Jets at higher energies are boosted more strongly, resulting in smaller average distances between particles in the jet. This results in a worse separation of particle inside the jet, and thus a worse resolution. Figure 22.6 shows an event display of a simulated hadronic jet in the ILD detector concept for the ILC with particle flow objects reconstructed by PandoraPFA. The benefit of a highly granular detector system is clearly visible.

Over the last 10 years, the Pandora algorithm has matured into a robust and stable algorithm. It is now used not only in the linear collider community, but also in long baseline neutrino experiments, and is under study at the LHC experiments.

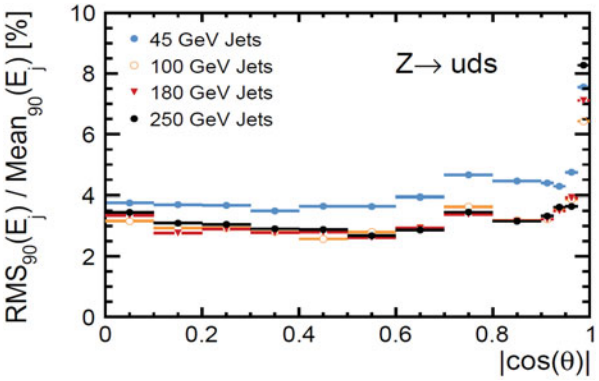


Fig. 22.5 The jet energy resolution, α, as a function of the dip angle $|\cos\theta_q|$ for jets of energies from 45 GeV to 250 GeV

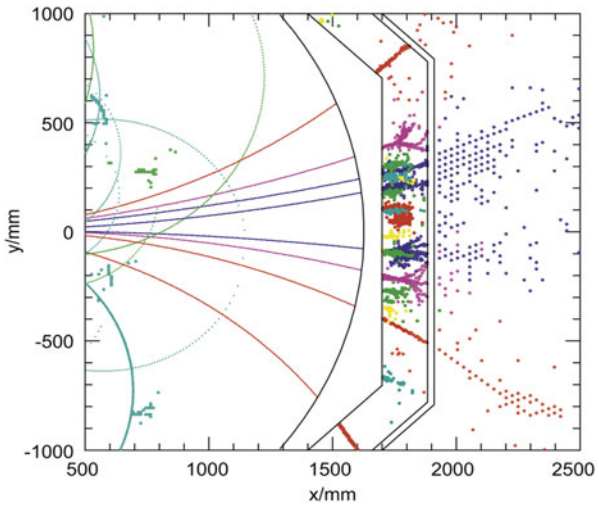


Fig. 22.6 Simulated jet in the ILD detector, with particle flow objects reconstructed by the Pandora algorithm shown in different colors

22.5.2 A Detector Concept for a Lepton Collider

Over the years a number of concepts for integrated detectors have been developed for use at a lepton collider [32–36]. Broadly speaking two different models exist: one based on the assumption, that particle flow is the optimal reconstruction technique, the other not based on this assumption. Common to all proposals is that both the tracking system and the calorimeter systems are placed insides a large superconducting coil which produces a large magnetic field, of typically 3–5 T. Both concepts use high precision tracking and vertexing systems, inside solenoidal

fields, which are based on state of the art technologies, and which really push the precision in the reconstruction of the track momenta and secondary vertices. Differences exist in detail in the choice of technology for the tracking devices, some rely heavily on silicon sensors, like the LHC detectors, others propose a mixture of silicon and gaseous tracking. The calorimeters are where these detectors are most different from current large detectors. The detectors based on the particle flow paradigm propose calorimeters which are more like very large tracking systems, with material intentionally introduced between the different layers. Systems of very high granularity are proposed, which promise to deliver unprecedented pictures of showering particles. Another approach is based on a more traditional pure calorimetric approach, but on a novel technology which promises to eventually allow the operation of an effectively compensated calorimeter [34].

At the ILC, detectors optimised for particle flow have been chosen as the baseline. The two proposed detector concepts ILD [32] and SiD [33] differ in the choice of technology for the tracking detectors, and on the overall emphasis based on particle flow performance at higher energies. However, both detectors have been optimised for collision energies of less than 1 TeV, while within the CLIC study the detector concepts have been further evolved to be optimised for operation at energies up to 3 TeV [35].

A conceptual picture of the ILD detector, as proposed for the ILC, is shown in Fig. 22.7. Visible are the inner tracking system, the calorimeter system inside the coil, the large coil itself, and the iron return yoke instrumented to serve as a muon identification system. A cut view of a quadrant with the sub-systems of ILD is shown in Fig. 22.8.

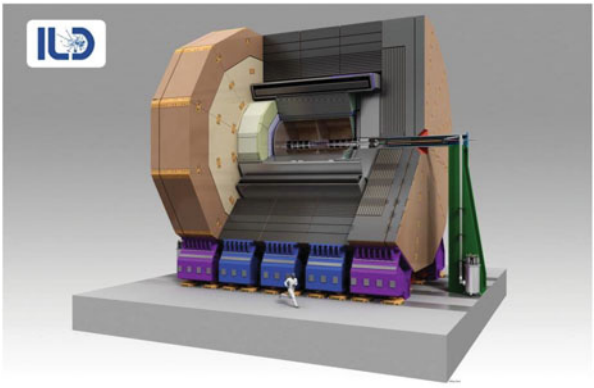


Fig. 22.7 Three-dimensional view of a proposed detector concept for the ILC, the ILD detector [32] (credit Ray Hori, KEK)

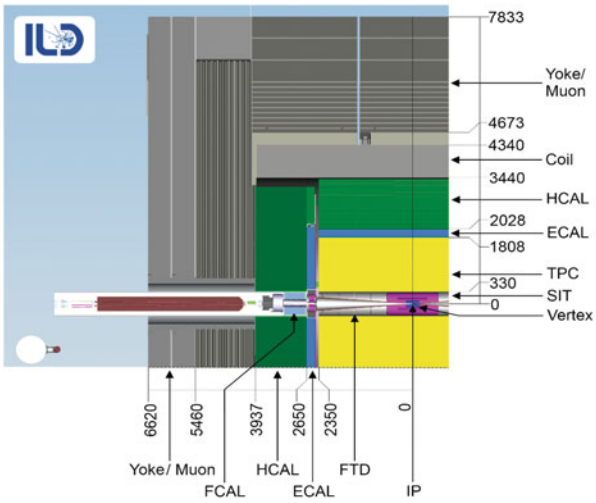


Fig. 22.8 Cut through the ILD detector in the beam plane, showing one quarter of the detector [37]

22.6 Detector Subsystems

A collider detector has a number of distinct sub-systems, which serve specific needs. In the following the main systems are reviewed, with brief descriptions of both the technological possibilities, and the performance of the system.

22.6.1 Trends in Detector Developments

Detector technologies are rapidly evolving, partially driven by industrial trends, partially itself driving technological developments. New technologies come into use, and disappear again, or become accepted and well-used tools in the community. A challenge for the whole community is that technological trends change faster than ever, while the design, construction and operation cycles of experiments become longer. Choosing a technology for a detector therefore implies not only using the very best available technology, but also one which promises to live on during the expected lifetime and operational period of the experiment. An example of this are Silicon technologies, which are very much driven by the demands of the modern consumer electronics industry. By the times Si detectors are operational inside an experiment, the technology used to built them is often already outdated, and replacements or extensions in the same technology are difficult to get. Even more so than to the sensors this applies to readout and data acquisition systems.

Because of the rapid progress in semiconductor technology, feature sizes in all kind of detector are getting ever smaller. Highly integrated circuits allow the integration of a great deal of functionality into small pixels, allowing the pixellation of previously unthinkable volumes. This has several consequences: the information about an event, a particle, a track, becomes ever larger, with more and more details at least potentially available and recorded. More and more the detection of particles, of properties of particles, rely no longer on averaging its behaviour over a volume large compared to the typical distances involved in the process used to measure the particle, but allows the experimenter to directly observe the processes which eventually lead to a signal in the detector. Examples of this are e.g. the Si-TPC (Silicon readout Time Projection Chamber, described in more detail below) where details of the ionisation process of a charged particle traversing a gas volume can be observed, or the calorimeter readout with Si-based pixellated detectors, given unprecedented insights into the development of particle showers. Once the volume read out becomes small compared to the typical distances involved in the process which is being observed, a digital readout of the information can be contemplated. Here, only the density of pixels is recorded, that is, per pixel only the information whether or not a hit has occurred, is saved. This results in potentially a much simpler readout electronics, and in more stable and simpler systems. These digital approaches are being pursued by detectors as different as a TPC and a calorimeter.

Increasing readout speed is another major direction of developments. It is coupled but not identical to the previously discussed issue of smaller and smaller feature sizes of detectors. Because of the large number of channels, faster readout systems need to be developed. An even more stringent demand however comes from the accelerators proposed, and the luminosities needed to make the intended experiments. They can only be used if data are readout very quickly, and stored for future use. To give a specific example: the detector with the largest numbers of pixels ever built so far (until the Phase-II upgrades of the LHC detectors) has been the SLD detector at SLAC which operated during the 1990. The vertex detector, realised from charged coupled sensors with some 400 Million channels, was readout with a rate of around 1 MHz. For the ILC readout speeds of at least 50 MHz, maybe even more, are considered, to cope with larger data rates and smaller inter-bunch spacings.

Technological advances in recent years have made it feasible to consider the possibility to do precision timing measurements with semi-conductor detectors. Timing resolutions in the range of 100 ps or better are becoming feasible, something completely unthinkable only a few years ago. This capability—somewhat orthogonal to the readout speed discussed above—can significantly extend the capabilities of semiconductor tracker, into the direction of so-called 4D tracking or calorimeter systems. Timing information at this level of precision can be used to measure the mass of particles through time-of-flight, and can help to separate out-of-time background from collision related events.

For many applications, particularly at the LHC, radiation hardness is at a premium. Major progress has been made in recent years in understanding damage mechanisms, an understanding, which can help to design better and more radiation hard detectors. For extreme conditions novel materials are under investigation.

22.6.2 Vertex Detectors: Advanced Pixel Detectors

Many signals for interesting physics events include a long lived particle, like e.g. a B or charmed hadron, with typical flight distances in the detector from a few 10 μm to a few mm. The reconstruction of the decay vertices of these particles is important to identify them and to distinguish their decay products from particles coming from the primary vertex, or to reconstruct other event quantities like vertex charge.

To optimally perform these functions the vertex detector has to provide high precision space points as close as possible to the interaction point, has to provide enough space points, so that an efficient vertex reconstruction is possible over the most relevant range of decay distances, of up to a few cm in radius, and present a minimal amount of material to the particle so as to not disturb their flight path. Ideally, the vertex detector also offers enough information that stand-alone tracking is possible based only on vertex detector hits.

At the same time a vertex detector has to operate stably in the beam environment. At a hadron collider it has to stand huge background rates, and cope with multiple interactions. At a lepton collider, very close to the interaction point a significant number of beam background particles may traverse the detector, mostly originating from the beam–beam interaction. These background particles are bent forward by the magnetic field in the detector. The energy carried away by this beamstrahlung may be several 10 TeV, which, if absorbed by the detector, would immediately destroy the device. The exact design of the vertex detector therefore has to take into account these potential backgrounds. At a hadron collider, the largest challenge will be to design the detector such that it can survive the radiation dose and is fast enough and has small enough pixels to cope with the large particle multiplicity. Here pixel size, readout speed, and radius of the detector are the main parameters which need to be optimised. At a lepton collider, both size and magnetic field can be used to make sure that the detector stays clear of the majority of the background particles. The occupancy at any conceivable luminosity is not driven by the physics rate, but only by the background events. Since they are much softer than the physics events, a strong magnetic field can be used to reduce the background rates, and allow small inner radii of the system. Nevertheless, the remaining hits from beam background particles dominate the occupancies, especially at the innermost layers of a vertex detector, and therefore require fast read-out speeds.

The particular time structure of the collider has an important impact on the design and the choice of the technology. At the ILC collisions will happen about every 300 ns to 500 ns, in a train of about 1 ms length, followed by a pause of around 200 ms. About 1300 bunches are expected to be in one train. A fast readout of the vertex detector is essential to ensure that only a small number of bunches are superimposed within the occupancy of the vertex detector. At CLIC the inter-bunch spacing is much smaller, putting a premium on readout speed. At LHC the typical time between collisions in a bunch crossing is order 100 ps decreasing to about 10 ps at the high luminosity LHC-HL.

A Si-pixel based technology is considered the only currently available technology which can meet all these requirements. A small pixel size ($< 20 \times 20\,\mu m^2$) combined with a fast read out will ensure that the occupancy due to backgrounds and from expected signals together remain small enough to not present serious reconstruction problems. It also allows for a high space point resolution, and a true three dimensional reconstruction of tracks and vertices essentially without ambiguities. Several silicon technologies are available to meet the demands. Increasingly, sensors based on the CMOS process are considered. Most recently devices with intrinsic gain larger than one are studied intensely, as they promise excellent performance combined with very good timing properties.

Quite a number of different technologies are currently under study. Broadly they can be grouped into at least two categories: those which try to read the information content as quickly as possible, and those which try to store information on the chip, and which are readout during the much longer inter-bunch time window. Another option under study is a detector with very small pixels, increasing the number of pixels to a point where even after integration over one full bunch train the overall occupancy is still small enough to allow efficient tracking and vertexing.

A fairly mature technology is the CCD technology [38, 39], which for the first time was very successfully used at the SLD experiment at the SLC collider at SLAC, Stanford. Over the past decade a number of systems based on this concept have been developed.

Newer approaches use the industrial CMOS process to develop monolithic active pixel sensors (MAPS) that are at the same time thin, fast, and radiation hard enough for particle physics experiments [40]. A smaller scale application of this technology is a series of test-beam telescopes, based on the Mimosa families of chips [41], built under the EUDET and AIDA European programs [42, 43] and operated a CERN, DESY and SLAC. The Phase-II upgrade of the ALICE experiment at the LHC contains a new inner tracking system that will be completely based on the CMOS-MAP sensor ALPIDE [44]. With a pixel size of $24.9\,\mu m \times 29.3\,\mu m$, a spatial resolution of $\approx 5\,\mu m$ and a time resolution of 5–10 µs is envisaged for hit rates of about $10^6/cm^2/s$. The CBM experiment, planned for the FAIR heavy-ion facility in Darmstadt, foresees to use MAPS for the microvertex detector. It will be based on the MIMOSIS chip, that is an advancement of the ALPIDE chip with similar pixel size and spatial resolution, but that has to cope with a much higher event rate of about $10^8/cm^2/s$ (and the associated radiation load) at the cost of a higher power consumption. The MIMOSIS chip already aims for a higher readout speed of about 5 µs.

In Fig. 22.9 a measured point resolution achieved with the CMOS-MAPS technology in a test beam experiment is shown [49]. Other technologies are at a similar level of testing and verifying individual sensors for basic performance.

Studies are underway to push the CMOS-MAPS towards even higher readout speeds [45]. The two parameters that currently govern the process are the time required for the pixel address encoding and the signal shaping during the pre-amplification. Changing the algorithm of the pixel address encoding and increasing the internal clock, could lead to an improvement from 50 ns to 25 ns for this step.

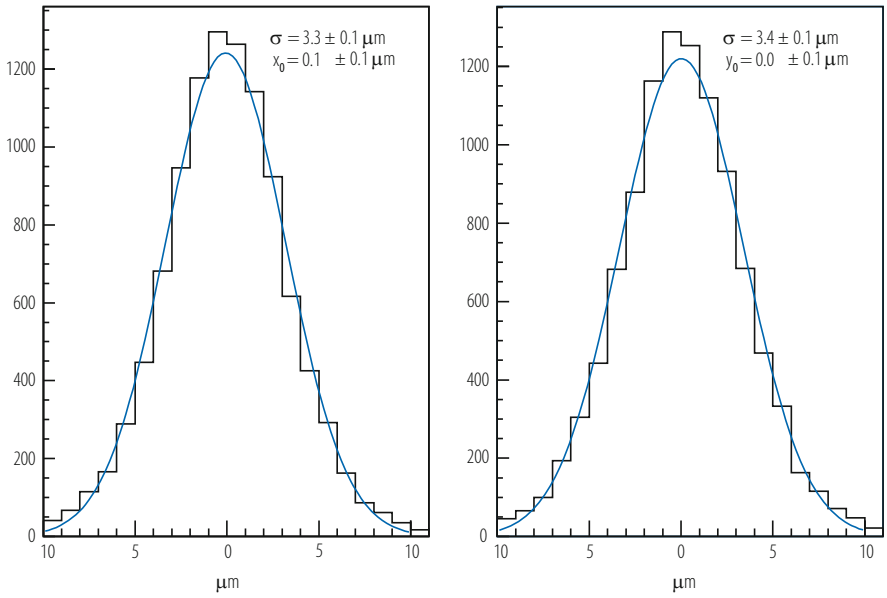


Fig. 22.9 (Left): Biased residual distribution measured in a CMOS pixel detector with 6 GeV electrons. (Right) Measured residual width in a 6 layer setup with a layer spacing of 20 mm [49]

The signal shaping currently takes about 2 μs and could be shortened to about 500 ns at the price of increasing the pixel current and therefore also increasing the power consumption. However, as the detectors at a linear collider would be operated in power-pulsing mode, the impact on the cooling requirements would be minor. Such an optimised CMOS detector for the ILC would have a readout speed of about 1 μs, i.e. it could be read out every two to three bunch crossings. Other groups explore the possibility to store charge locally on the pixel, by including storage capacitors on the pixel. Up to 20 timestamped charges are foreseen to be stored, which will then be readout in between bunch trains.

The most recent example of a pixel detector at a lepton collider has been the pixel detector for the Belle-II experiment. This system is based on the DEPFET technology [46]. Charge generated by the passage of a charged particle through the fully depleted sensitive layer is collected on the gate of a DEPFET transistor, implemented into each pixel. DEPFET sensors can be thinned to remove all silicon not needed for charge collection, to something like 50 μm, or 0.1% of a radiation length. This makes this technology well suited for lepton collider applications, where minimal material is of paramount importance [48].

A problem common to all technologies considered is the amount of material present in the detector. A large international R&D program is under way to reduce significantly the material needed to build a self-supporting detector. The goal, driven by numerous physics studies, and the desire for ultimate vertex reconstruction, is a single layer of the detector which in total presents 0.1% of a radiation length,

including sensor, readout and support. This can only be achieved by making the sensors thin, and by building state-of-the-art thin and light weight support structures. To compare, at the LHC the total amount of material present in the silicon based trackers is close to 2 radiation length, implying that per layers, close to 10% of a radiation length is present.

Very thin sensor layers are possible with technologies based on fully depleted sensors. Since here only a thin layer of the silicon is actually needed for the charge collection the rest of the wafer can be removed, and the sensor can be thinned from typically 300 µm, used e.g. in the LHC experiments, to something like 50 µm or less. Several options are under study how such thin Si-ladders can then be supported. One designs foresees that the ladders be stretched between the two endcaps of the detectors, being essentially in the active area without additional support. Another approach is to study the use of foam material to build up a mechanically stiff support structure. Carbon foam is a prime candidate for such a design, and first prototype ladders have come close to the goal of a few 0.1% X_0 [47]. Another group is investigating whether Si itself could be used to provide the needed stability to the ladder. By a sophisticated etching procedure stiffening ribs are built into the detector, in the process of removing the material from the backside, which will then stabilise the assembly. This approach has been successfully implemented for the vertex detector at the Belle-II experiment [48].

Material reduction is an area where close connections exist between developments done for the ILC and developments done for the LHC and its upgrade. In both cases minimum material is desired, and technologies developed in the last few years for the ultra-low material ILC detector are of large interest to possible upgrade detectors for LHC and LHC Phase-II.

The readout of these large pixel detectors present in itself a significant challenge. On-chip zero-suppression is essential, but also well established. Low power is another important requirement, consistent with the low mass requirement discussed above. Only a low power detector can be operated without liquid cooling, low mass can only be achieved without liquid cooling. It has been estimated that the complete vertex detector of an ILC detector should not consume on average more than 100 W, if it is to be cooled only through a gas cooling system. Currently this is only achievable if the readout electronics located on the detector is switched off for a good part of the time, possible with the planned bunch structure of the ILC. However such a large system with pulsed power has never been built, and will require significant development work. It should not be forgotten that the system needs to be able to operate in a large magnetic field, of typically 4 T. Each switching process therefore, which is connected with large current flows in the system, will result in large Lorentz forces on the current leads and the detectors, which will significantly complicate the mechanical design of the system. Nevertheless, with current technologies power pulsing is the only realistic option to achieve the desired low power operation, and thus a central requirement for the low mass design of the detector. In Fig. 22.10 the conceptual layout of a high precision vertex detector is shown.

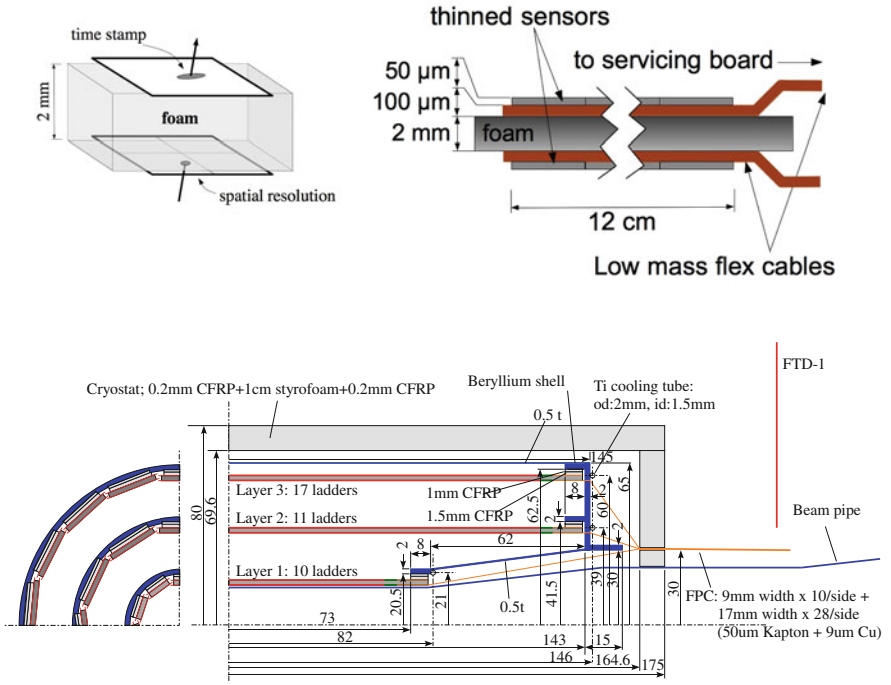


Fig. 22.10 Top: Concept of a double-layer vertex detector system developed within the PLUME project. Bottom: Vertex detector for the ILD concept, based on a layout with three double layers [37]

One of the key performance figures of a vertex detector is its capability to tag heavy flavours. At the ILC b-quarks are an important signature in many final states, but more challenging are charm quarks as the are e.g. expected in decays of the Higgs boson. Obtaining a clean sample of charm hadrons in the presence of background from bottom and light flavour is particularly difficult. Already at the SLC and the LEP collider, the ZVTOP [50] algorithm has been developed and used successfully. It is based on a topological approach to find displaced vertices. Most tracks originating from heavy flavour decays have relatively low momenta, so excellent impact parameter resolution down to small ($\approx$1GeV) energies is essential. On the other hand, due to the large initial boost of the heavy hadrons, the vertices can be displaced by large distances, up to a few cm away from the primary vertex, indicating that the detector must be able to reconstruct decay vertices also at large distances from the interaction point. The algorithms have been further developed and adapted to the expected conditions at a linear collider [51]. The performance of a typical implementation of such a topological vertex finder is shown in Fig. 22.11.

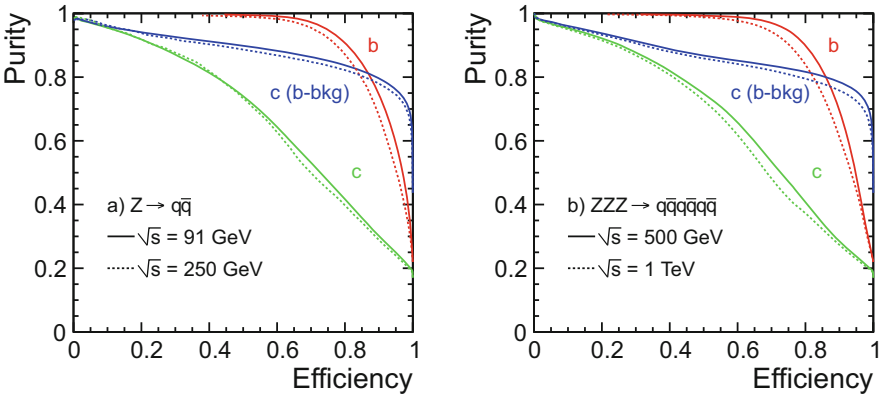


Fig. 22.11 Purity versus efficiency curve for tagging b-quarks (red points) and c-quarks (green points) and c-quarks with only b-quark background (blue points) obtained in a simulation study for Z-decays into two (left) and six (right) jets, as simulated in the ILD detector [37]

22.6.3 Solid State Tracking Detectors: Strip Detectors

To determine the momentum of a charged particle with sufficient accuracy, a large volume of the detector needs to be instrumented with high precision tracking devices, so that tracks can be reliably found and their curvature in the magnetic field can be well measured. Cost and complexity considerations make a pixel detector for such large scale tracking applications at present not feasible. Instead strip detectors are under development, which will provide excellent precision in a plane perpendicular to the electron-positron beam.

Silicon microstrip detectors are extremely well understood devices, which have been used in large quantities in many experiments, most recently on an unprecedented scale by the LHC experiments. A typical detector fabricated with currently established technology might consist of a 300 µm thick layer of high resistivity silicon, with strips typically every 50 µm running the length of the detector. Charge is collected by the strips. These detectors measure one coordinate very well, with a precision of <10 µm. The second coordinate can be measured e.g. by arranging a second layer of strip detectors at a small stereo angle. Double sided detectors, with two readout structures on either side, with strips running also at an angle to each other, have in the past proved to be a costly and not very reliable alternative to the combination of two single sided detectors back-to-back.

Strip detector have received a major boost through the upgrade program for the LHC experiments. The large area tracking systems for both ATLAS and CMS will need to be replaced in time for the start of the high luminosity phase of the LHC, scheduled to start around 2026. Several hundred square meters of Silicon detectors need to be produced, to build up these large detector systems. Compared to the previous ones, the radiation hardness of these devices had to be improved by at least an order of magnitude, and the total amount of material in the system will be reduced

significantly. This requires novel approaches to the structures, and to powering and cooling of these detectors, which will be discussed in a separate section.

A major R&D goal needed for the application of these devices to the ILC detector is the significant reduction of material per layer. As for the vertex pixel detector, thinning the detectors is under investigation, as is the combination of thinned detectors with light weight support structures and power-pulsed readout electronics. New schemes to deliver power to the detectors—like serial powering—are being studied.

22.6.4 Gaseous Tracking

Even though solid state tracking devices have advanced enormously over the last 20 years, gaseous tracking is still an attractive option for a high precision detector like an ILC detector. Earlier in this section the concept of particle flow has been discussed. Particle flow requires not the very best in precision from a tracking detector, but ultimate efficiency and pattern recognition ability. Only if charged tracks are found with excellent efficiency can the concept of particle flow really work. A large volume gaseous tracker can assist in this greatly by providing a large number of well measured points along a track, over a large volume. In addition a gaseous detector can assist in the identification of the particle by measuring the specific energy loss, dE/dx, of the particle, which for moderate momenta up to 10–20 GeV is correlated to the particle type.

A particularly well suited technology for this is the time projection chamber, TPC [52]. It has been used in the past very successfully in a number of colliding beam experiments, most recently in the ALICE experiment at the LHC [53]. A time projection chamber (see Chapt. C1 ii) essentially consists of a volume of gas, onto which a uniform electric and magnetic field is superimposed. If a charged particle crosses the volume, the produced ionisation is drifted under the influence of the field to the anode and the cathode of the volume. Since the electrons drift typically about 1000 times faster than the ions, they are usually used in the detection. A gas amplification system at the anode side is used to increase the available charge which is then detected on a segmented anode plane, together with the time of arrival. Combining both, a three dimensional reconstruction of the original crossing point is possible.

Traditionally time projection chambers are read out at the anode with multi-wire proportional chambers. They operate reliably, have a good and well controllable gas gain, and give large and stable signals. However wires are intrinsically one dimensional, which means, that a true three-dimensional reconstruction of the space point is difficult. Wires need to be mechanically stretched, which restricts the distance between them to something larger than typically 1 mm. More importantly though, the fact that all electrons produced in the drift volume are eventually collected by these wires, and that this collection happens in a strong magnetic field, limits the achievable resolution. Very close to the wire the electric field lines and the

magnetic field lines are no longer parallel, and the particle will start to deviate from the ideal straight track toward the anode. It will start to see a strong Lorentz force, which will tend to distort the drift path. This distortions will be different whether the electron approaches the wire from below or from above, and will introduce biases in the reconstruction of the space coordinate which might be similar in size to the spacing between the wires. Corrections might be applied, and can correct in part this effect, but typical uncertainties around 1/10 of the inter-wire distance might remain. This does limit the ultimately achievable resolution in a wire-equipped TPC.

An alternative which is being studied intensely is the use of micro-pattern gas detectors as readout systems in a TPC [54]. Gas electron multipliers (GEM) [55, 56] or Micro Mesh Chambers Micromegas (MM) [57, 58] are two recent technologies under investigation.

A GEM foil consists of a Polyamide foil of a typical thickness of 50–100 μm, copper clad on both sides. A regular grid of holes of 50 μm diameter spaced typically 150 μm apart connects the top and the bottom side. With a potential of a few hundred volts applied across the foil a very high field develops inside the hole, large enough for gas amplification. Gains in excess of 10^3 have been achieved with such setups. In Fig. 22.12 the cross section of a hole in a GEM is shown, together with field lines, showing clearly the high field region in the center of the hole. A challenge for the GEM based system is the development of a mechanically stable readout system. A system based on ceramic spacer structures has been developed and successfully tested [61].

A MM is constructed by stretching a metal mesh with a very fine mesh size across a readout plane, at a distance of typically less than 1 mm from the readout plane. A potential is applied between the mesh and the readout plane. The resulting field is large enough for gas amplification. Spacers at regular intervals ensure that the system is mechanically stable, and withstands the electrostatic forces.

Both systems have feature sizes which are one order of magnitude smaller than the ones in conventional wire-based readout systems, thus reducing the potential errors introduced through the gas amplification system. The smaller feature sizes in addition reduce the spatial and temporal spread of the signals arriving at the readout structure, thus promising a better two particle separation. The spatial resolution obtained in a prototype TPC equipped with a Micromegas readout is shown in Fig. 22.13.

The positive ions which are produced both in the initial ionisation along the track, and in the amplification process at the anode, will drift slowly to the cathode. Thus, the drift volume of the TPC will slowly fill with positive charge, if nothing is done, which will tend to change the space-to-time relation central to the TPC principle. Both GEM and MM suppress the drift of positive ions to the cathode, by catching a large percentage (over 98%) on the GEM foil or on the mesh [60]. To reduce the amount of positive ions even further a gating electrode can be considered. This is an electrode mounted on top of the last amplification stage, facing towards the drift volume. The potential across the gate can be changed to change the transparency of the gate for ions. At the ILC the gate can be opened for one complete bunch train, and then be closed for the inter-bunch time. This would reduce the volume affected by significant ion densities to only the first few cm in drift, above the readout plane.

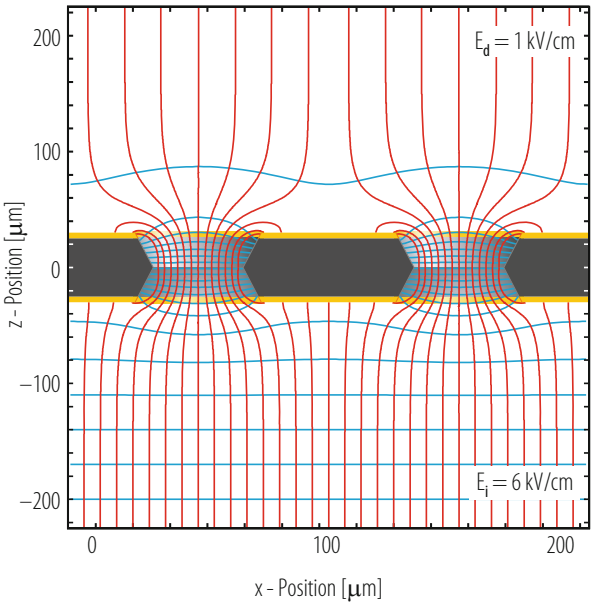


Fig. 22.12 Cross section of a hole in a GEM foil, with simulated field lines (picture credit Oliver Schäfer, DESY)

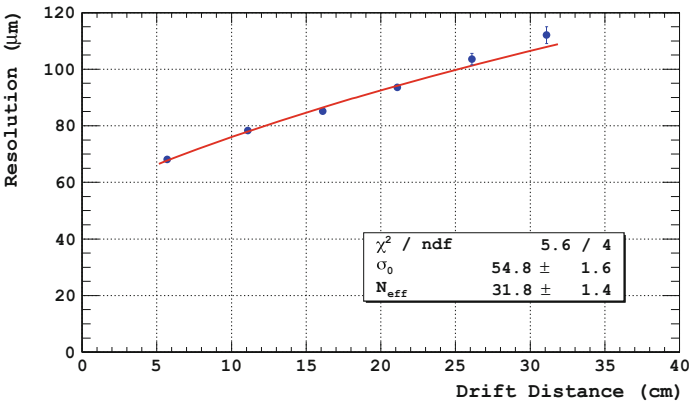


Fig. 22.13 Preliminary result of the spatial resolution of Micromegas readout as a function of drift length. A resistive pad plane was used to spread the charge [59]

Recently specialised GEM foils have been developed, which show a very large optical transparency. Experimentally is has been shown that such devices allow a large change in electron transparency, from close to 90% to 0%, by changing the potential across the GEM by some 50 V. This is expected to translate into a very similar change in ion-transparency, but the final experimental proof for this is still missing.

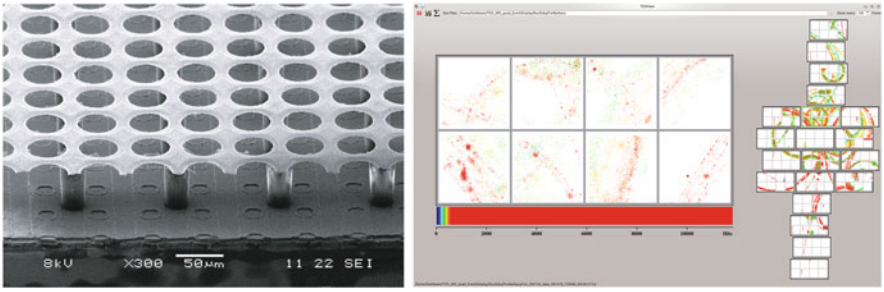


Fig. 22.14 Left: Microscopic picture of an Ingrid: a micromegas detector implemented on top of the read out chip by post-processing; Right: Event display of test beam electrons in a Pixel-TPC setup with Ingrids and Micromegas readout [62] (Credit Michael Lupberger, Bonn)

A recent development tries to combine the advantages of a micro-pattern gas detector with the extreme segmentation possible from silicon detectors. A Si pixel detector is placed at the position of the readout pad plane, and is used to collect the charge behind the gas amplification system. Each pixel of the readout detector has a charge sensitive amplifier integrated, and measures the time of arrival of the signal. Such a chip was originally developed for medical applications (MediPIX [63]), without timing capability, and has since been further developed to also include the possibility to record the time (Timepix [64]). This technology, which is still in its infancy, promises exciting further developments of the TPC concept. The close integration of readout pad and readout electronics into one pixel allows for much more compact readout systems, and also for much smaller readout pads. Pad sizes as small as $50 \times 50\,\mu$ m have been realised already. This allows a detailed reconstruction of the microscopic track a particle leaves in the TPC, down to the level of individual ionisation clusters. First studies indicate that a significantly improved spatial resolution can be obtained through silicon pixel readout of the TPC. In Fig. 22.14 a picture of a track segment recorded in a small test setup equipped with a Micromegas and the Medipix chip is shown.

The size of charge clouds in a typical TPC is of the order of a few hundred µm to mm, depending on the choice of gas, on environmental parameters like pressure and magnetic field, and on the drift distance. The feature size of the proposed silicon based readout is significantly smaller than this, which may allow the operation of the TPC in a different mode, the so called digital TPC mode. In this case no analogue information about the size of the charge collected at the anode is recorded, but instead only the number and the distribution of pixels which have fired are saved. The distribution of the hits is used to reconstruct the position of the original particle, much as it is done in the case of a conventional TPC. It can be shown that as long as the pixel size is small compared to the size of the electron cloud the number of pixels is a good measure for both the position of the cluster and the total charge in the cluster. One advantage of recording only the number of hits is that the sensitivity to delta rays is reduced. Delta rays are energetic electrons which are kicked out of a

gas molecule by the interaction with the incoming particle, and which then rapidly loose energy in the gas. Delta rays produce large charge clusters along the track, which are not correlated any more with the original particle. They also produce charge some distance away from the original track, and thus limit the intrinsic spatial resolution. Altogether delta rays are responsible for the tails in the charge distribution along a particle track, and for a deterioration of the possible spatial resolution. In digital readout mode these effects are less pronounced. The tails in the charge distribution are reduced, and the excellent spatial resolution through small pads allows the removal of at least some delta rays on a topological basis. Recent studies indicate that the spatial resolution of a Si based TPC readout might be better by about 30%, while the capability to measure the specific energy loss, dE/dx, might increase by 10–20% [65].

22.6.5 Electromagnetic Calorimeters

The concept of particle flow discussed above requires an excellent granularity in the calorimeters to separate charged from neutral particles in the calorimeter. Some hypothetical New Physics scenarios are associated with event topologies where high energetic photons do not originate at the interaction region, so that the device should in addition be able to also reconstruct the direction of a photon shower with reasonable accuracy.

Electromagnetic calorimeters (ECAL) are designed as compact and fine-grained sandwich calorimeters optimised for the reconstruction of photons and electrons and for separating them from depositions of hadrons. Sandwich calorimeters are the devices of choice, since they give information on the development of the cluster both along and transverse to the direction of the shower development. This capability is very difficult to realise with other technologies, and is essential to obtain an excellent spatial reconstruction of the shower. To keep the Moliére radius small, tungsten or lead are used as absorber. Sensor planes are made of silicon pad diodes, Monolithic Active Pixel sensors (MAPS) or of scintillator strips or tiles.

A major problem of fine-grained calorimeters is one of readout and data volume. For a typical electromagnetic calorimeter considered for the ILC, where cell sizes of $5 \times 5\,\mathrm{mm}^2$ are investigated, the number of channels quickly passes the million. With the progress in highly integrated electronics, more and more of the readout electronic is going to be integrated very close to the front-end. The design of the electromagnetic calorimeter by the CALICE collaboration [66] or by a North-American consortium [67, 68] has the silicon readout pads integrated into a readout board which sits in between the absorber plates. A special chip reads out a number of pads. A 12-bit ADC is included on the chip, and data are then sent on thin Kapton tape cables to the end of the module. There data from the different chips are concentrated, and sent on to the central data acquisition system. Such highly integrated detector designs have been successfully tested in large scale prototypes

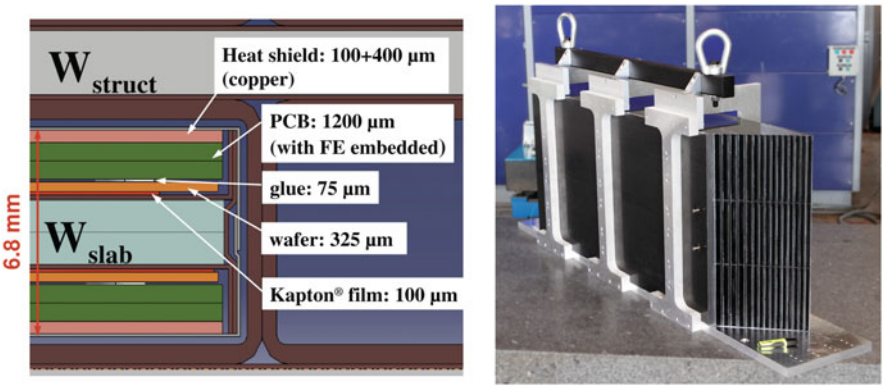


Fig. 22.15 Schematic figure of an integrated silicon-tungsten layer for an ILC ECAL (left) and tungsten absorber prototype (right) [37]

in test beams at CERN and Fermilab, although with a earlier version of the readout electronics, with a lesser degree of concentration (Fig. 22.15).

It is only with the progress in integration and in the resulting price reduction per channel that large scale Si-based calorimeter systems will become a possibility. Nevertheless the price for a large electromagnetic calorimeter of this type is still rather high, and will be one of the most expensive items in a detector for a linear collider. A cheaper alternative investigated is a more conventional sampling calorimeter readout by Scintillator strips. Two layers of strips at orthogonal orientation followed by a somewhat larger tile can be used to result in an effective granularity as small as $1 \times 1\,\mathrm{cm}^2$, nearly as good as in the case of the Si-W calorimeter. Light from the strips and tiles is detected from novel silicon based photo-detectors (for a more detailed description, see the section on hadronic calorimeters). The reconstruction of the spatial extent of a shower in such a system is more complicated, since ambiguities arise from the combination of the different layers. In addition the longitudinal information of the shower development is less detailed, but still superb compared to any existing device. This technology as well has been successfully tested in test beam experiments, and has shown its large potential.

Whether or not this technology or the more expensive Si-W technology is chosen for a particular detector depends on the anticipated physics case, and also the center-of-mass energy, at which the experiment will be performed. Simulation studies have shown that at moderate energies, below 250 GeV, both technologies perform nearly equally well, only at larger energies does the more granular solution gain an advantage. To some extent this advantage can be compensated by a larger detector in the case of the scintillator, though the price advantage then quickly disappears.

An extreme ansatz is a study trying to use vertex detector technology as readout planes in a calorimeter. The MAPS technology has been used to equip a tungsten absorber stack with sensors. This results in a extremely fine granular readout, where again only digital information is used—that is, only the number of pixels hit within

a certain volume is used, not any analogue information. This in turn means a much simpler readout electronics per channel, and a potentially more robust system against noise and other electrical problems. The amount of detail which can be reconstructed with such a system is staggering, and would open a whole new realm of shower reconstruction. However the cost at the moment is prohibitive, and many technical problems would need to be solved should such a system be used on a large scale [69].

22.6.6 Hadronic Calorimeters

In a particle flow based detector the distinction between an electromagnetic and a hadronic calorimeter conceptually disappears. Finely grained systems are needed to reconstruct the topology of the shower, both for electromagnetically and for hadronically interacting particles. Nevertheless, the optimization of the hadronic section of the calorimeter results in a coarser segmentation.

The traditional approach is based on a sampling calorimeter, typically with iron as absorber, maybe with lead, and with scintillators as active medium. New semiconductor photo detectors allow the individual readout of comparatively small scintillator tiles. These photo detectors are pixellated Si diodes, with of order 1000 diodes on an area of 1 mm^2. Each diode is operated in the limited Geiger mode, and the number of photons detected is read out by counting the number of pixels which have fired. This is another example of the previously discussed digital readout schemes. These so-called silicon photo multipliers (SiPM), also called Multi Pixel Photon Counters (MPPC), are small enough that they can be integrated into a calorimeter tile. To operate they only need to be provided with a potential of below 100 V, and the power lines are used to read out the signal from the counter. This makes for a rather simple system, which allows the instrumentation of a large number of tiles, and thus the construction of a highly granular scintillator based calorimeter. Complications which in the past severely limited the number of available channels—e.g., the routing of a large number of clear fibers from the tile to the photon counter, the operation of a larger number of bulky photo-multipliers of rather high voltage, etc all do not apply any more.

Light created through scintillation in the tile is collected by a Silicon photomultiplier, attached to each tile Earlier systems needed a wave-length shifting fibre, to adopt to the spectral sensitivity of the sensor (c.f. Fig. 22.16). A calibration of the energy response of the tile and SiPM system has two components: For small signals, the output signal shows contributions from one, two, three and more photons by clearly separate peaks in the amplitude spectrum. These can be used to establish the response of the system to single photons. At high signals, because of the limited number of pixels on the sensor, saturation leads to a non-linear response of the system. This needs to be measured and calibrated on the test bench, using a well calibrated photon source.

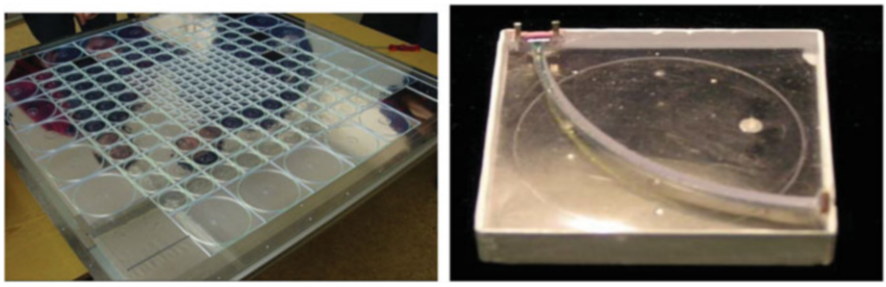


Fig. 22.16 Picture of a prototype readout plane for a highly segmented tile calorimeter (left) and one scintillator tile with wavelength shifting fibre and SiPM readout (right) [70]

The CALICE collaboration has designed a calorimeter based on this technology to be used in a detector at the linear collider. It is based on steel as absorber material, and uses $3 \times 3\,\mathrm{cm}^2$ scintillator tiles as sensitive elements. Each tile is readout by a silicon photomultiplier. A prototype readout plane is shown in Fig. 22.16. Groups of tiles are connected to a printed circuit board, which provides the voltage to the SiPM's, and routes the signals back to a central readout chip. This chip, which has been derived from the one developed for the Si-W calorimeter readout described in the previous section, digitises the signals, multiplexes them and sends them out to the data acquisition. Again, nearly all of the front-end electronics is integrated into this printed circuit board, and as such becomes part of the readout plane. This makes for a very compact design of the final calorimeter, with minimum dead space, and only a small number of external connections. This calorimeter has successfully passed a series of stringent beam tests in recent years, giving confidence that this technology is mature and can be used for a large scale detector application.

Recently the technology for SiPM advanced and pushed the sensitivity into the ultra-violet range, making a direct coupling between scintillator and silicon sensor possible (c.f. Fig. 22.17). The SiPM-on-tile technology has been proposed for the upgrade of the CMS endcap calorimeter. This system will use many of the developments done for an ILC detector, and be the first large-scale real-life application of this technology in an experiment. Through significantly smaller in size than the anticipated linear collider experiment, it will be a major asset for the LC community. Figure 22.17 shows a prototype readout HCAL plane using the SiPM-on-tile technology.

A potentially very interesting development in this area is again a digital version of such a calorimeter [72]. If the tile size can be made small enough - for hadronic showers this means a few $1 \times 1\,\mathrm{mm}^2$—a digital readout becomes possible. Counting the number of tiles belonging to a shower gives a good estimate of the showers energy. However scintillator tiles are difficult to built and read out for sizes this small—a major problems is the coupling between the light and the photo detector—so that a gaseous option is considered for this digital approach. Resistive plate chambers offer a cheap and well tested possibility to instrument large areas. They

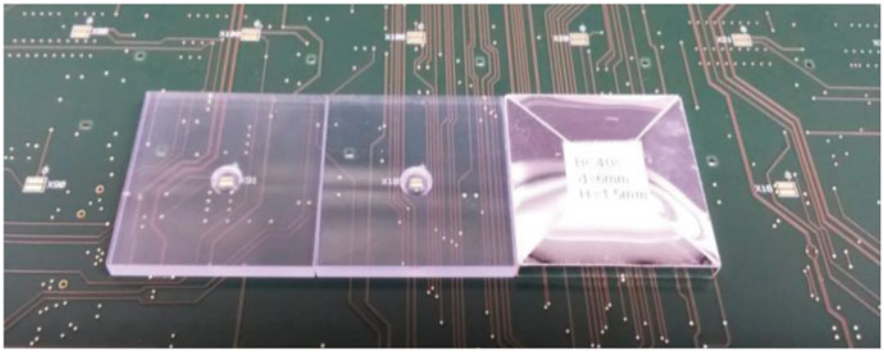


Fig. 22.17 Picture of HCAL scintillator tile with direct SiPM-on-tile readout [71]

are readout by segmented anode planes, which can be easily constructed with small pads of order of $1 \times 1\,\mathrm{cm}^2$. The principle of such a digital calorimeter has been established, and seems to meet specifications [72]. A major challenge however is to produce readout electronics for the very large number of channels which is about an order of magnitude cheaper per channel than the one for the analogue tile technology.

An interesting compromise between digital and analogue readout calorimeters is the semi-digital approach. Here, moderately dimensioned cell sizes of $1 \times 1\,\mathrm{cm}^2$ are combined with a rather simple 2-bit electronics with three signal thresholds. This would allow for having a high enough granularity to study the fine details of the hadronic shower evolutions and at the same time use the semi-digital charge signal for the analysis. A prototype semi-digital calorimeter for the ILD concept has been built and tested in beams and shows promising results [73].

A gaseous readout system has another feature which might be of advantage for a particle flow calorimeter. In the development of hadronic showers many neutrons are produced. Because of their long mean free path the loose energy and get absorbed far away from the core of the shower. This makes it very hard to attach these hits to the correct shower, thus creating a deficit in the energy for the shower, and creating fake hits away from the shower which might be confused with other nearby showers. Because of the very low cross section for neutrons in typical counter gases hardly any hits due to neutrons are recorded in a RPC based system. In a scintillator system, because of the high hydrogen and carbon content of the scintillator, the opposite is the case, and significant numbers if hits from neutrons are observed. On the other hand, neutrons travel slowly, and hits from neutron are later in time than other particle. Timing information at the 10 ns level might be good enough to reject a large number of the neutron hits in a shower. Its impact on the shower reconstruction is a subject of intense study at the moment, for both technologies, and no final verdict can be given which technology in the end has more advantages.

Large prototypes of ECAL and HCAL calorimeter systems have been built and tested in testbeam experiments. Figure 22.18 shows an event display for a combined

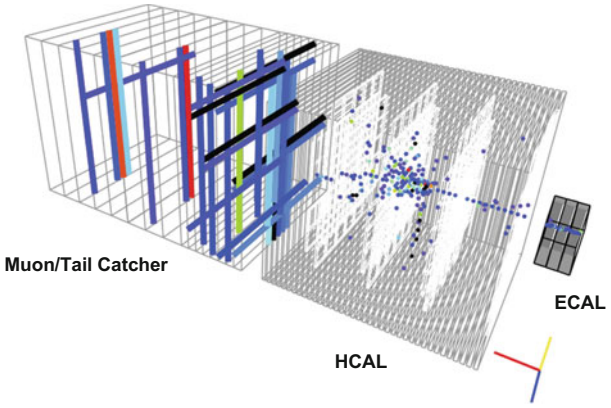


Fig. 22.18 Event in a combined testbeam where a 20 GeV pion (from the right) passes an ECAL prototype (small volume on the right), an analogue HCAL prototype with scintillator-tile readout (centre), and a muon system/tail catcher prototype with scintillator-strip readout (left) [71]

setup (from right to left) of a silicon-tungsten ECAL, an analogue scintillator-steel HCAL, and a muon/tailcatcher system with scintillator-strip readout (c.f. Sect. 22.6.7). A 20 GeV pion enters from the right, the details of the hadronic shower are clearly visible.

22.6.7 Muon Detectors

The flux return from the large field solenoids usually is realised as a thick iron return yoke. Often the iron is slit and detectors are integrated into the slots to serve as muon detectors. Many types of low-cost large-area charged particle detectors are possible and under investigation, e.g. resistive plate chambers, GEM chambers, or Scintillator based strip detectors. In a detector equipped with highly segmented calorimeters however a lot of the measurements traditionally done by such a muon system can be done in the calorimeters themselves. The identification of muons is greatly helped by the hadronic calorimeter, and its longitudinal sampling. Due to the high fields anticipated, muons below 3–4 GeV in fact never even reach the muon chambers, and need to be identified by the calorimeters together with the tracking system. The parameters of the muon are measured by the detector inside the coil, combining information from the tracker and the calorimeter. For these detector concepts the muon system in fact only plays a minor role, and can be used to backup and verify the performance of the calorimeter system.

An interesting approach is proposed by one of the ILC detector concepts [34]. Here the magnetic flux is returned not by an iron yoke, but by a second system of large coils. A smaller coil creates the high central field, of about 3 T, while a second larger coil creates a 1.5 T field in the opposite direction and serves as the flux return.

A system of planar coils in the endcap control the transition from the small to the large bore coil. In this concept muon chambers are mounted in between the two large solenoids. A similar approach is followed up in the studies for the very large detectors of potential very large hadron colliders.

22.6.8 Triggering at the ILC

The comparative cleanliness of events at the ILC allow for a radical change in philosophy compared to a detector at the LHC: the elimination of a traditional hardware based trigger. Triggering is a major concern at the LHC, and highly sophisticated and complex systems have been developed and built to reduce the very high event rate at the LHC to a manageable level [74, 75]. At the ILC with its clean events, without an underlying event, it is possible to operate the detector continuously and read out every bunch crossing. At a local level filtering is applied to the data to remove noise hits, and to eliminate as much as possible "bad hits", but overall no further data reduction is done. Events are written to the output stream unfiltered, and are only classified by software at a later stage. This allows the detector to be totally unbiased to any type of new physics, and to record events with the best possible efficiency. As a draw back the expected data rates are rather large. Great care has to be taken that the detector systems are robust and not dominated by noise, so that the data volume remains manageable, and the readout can keep up with the incoming data rate.

A slightly different approach has been suggested by the LHC experiments ALICE and LHCb, where upgrade plans foreseen to read out every event and to perform event selection and reconstruction in on-line processor farms.

22.7 Summary

Even though with the four LHC experiments, major experimental facilities recently built and commissioned, work on the next generation of experiments is proceeding. In particular the proposed linear collider poses very different and complementary challenges for a detector, with a strong emphasis on precision and details of the reconstruction. Significant work is happening worldwide on the preparation of technologies for this project. First results from test beam experiments show that many of performance goals are reachable or have already been reached. The move to ever increasing number of readout channels, with smaller and smaller feature sizes, has triggered a systematic investigation of "digital" detectors, where for a huge number of pixels only very little information per pixel is processed and stored. Whether or not such systems are really feasible in a large scale experiment is not proven yet. Tests over the next few years will answer many of these questions.

References

1. J. Gillies et. al (eds.), "Accelerating Science and Innovation", CERN-Brochure-2013-004-Eng. CERN 2013.
2. S. Ritz et. al, "Building for Discovery", Accessible at: http://www.usparticlephysics.org., 2014.
3. M. Nozaki et. al, "AsiaHEP/ACFA Statement on the ILC", Accessible at: http://www.acfa-forum.net/AsiaHEP/other_documents, 2014.
4. M. Nozaki et. al, "AsiaHEP/ACFA Statement on the ILC and CEPC/SPPC", Accessible at: http://www.acfa-forum.net/AsiaHEP/other_documents, 2016.
5. International Committee on Future Accelerators, "ICFA Statement on the ILC Operating at 250 GeV as a Higgs Factory", Accessible at: http://icfa.fnal.gov/statements, 2017.
6. T. Behnke *et al.*, "The International Linear Collider Technical Design Report - Volume 1: Executive Summary," arXiv:1306.6327 [physics.acc-ph].
7. M. Aicheler *et al.*, "A Multi-TeV Linear Collider Based on CLIC Technology : CLIC Conceptual Design Report," https://doi.org/10.5170/CERN-2012-007.
8. M. J. Boland *et al.* [CLIC and CLICdp Collaborations], "Updated baseline for a staged Compact Linear Collider," https://doi.org/10.5170/CERN-2016-004 arXiv:1608.07537 [physics.acc-ph].
9. M. Benedikt *et al.*, "Future Circular Collider : Vol. 2 The Lepton Collider (FCC-ee)," CERN-ACC-2018-0057.
10. M. Benedikt *et al.*, "Future Circular Collider : Vol. 3 The Hadron Collider (FCC-hh)," CERN-ACC-2018-0058.
11. [CEPC Study Group], "CEPC Conceptual Design Report: Volume 1 - Accelerator," arXiv:1809.00285 [physics.acc-ph].
12. CEPC-SPPC Study Group, "CEPC-SPPC Preliminary Conceptual Design Report." IHEP-CEPC-DR-2015-01, IHEP-TH-2015-01, IHEP-EP-2015-01.
13. M. Mangano, "Physics at the FCC-hh, a 100 TeV pp collider," CERN Yellow Report CERN 2017-003-M https://doi.org/10.23731/CYRM-2017-003 [arXiv:1710.06353 [hep-ph]].
14. CMS Collaboration, "Technical Proposal for a MIP Timing Detector in the CMS Experiment Phase 2 Upgrade," CERN-LHCC-2017-027.
15. I. Dawson [ATLAS and CMS Collaborations], "The SLHC prospects at ATLAS and CMS", J. Phys. Conf. Ser. **110** (2008) 092008.
16. J. Brau, Y. Okada, N. Walker (editors) [ILC Collaboration], "ILC Reference Design Report Volume 1 - Executive Summary", arXiv:0712.1950.
17. EUPRAXIA - European Plasma Research Accelerator with Excellence in Applications, see *http://www.eupraxia-project.eu*.
18. TESLA Technology Collaboration, see *http://tesla.desy.de*.
19. M. Altarelli *et al.*, "XFEL: The European X-Ray Free-Electron Laser. Technical design report," https://doi.org/10.3204/DESY_06-097
20. A. Grassellino *et al.*, "Nitrogen and argon doping of niobium for superconducting radio frequency cavities: a pathway to highly efficient accelerating structures," Supercond. Sci. Technol. **26** (2013) 102001 https://doi.org/10.1088/0953-2048/26/10/102001 [arXiv:1306.0288 [physics.acc-ph]].
21. C. Adolphsen *et al.*, "The International Linear Collider Technical Design Report - Volume 3.I: Accelerator R&D in the Technical Design Phase," arXiv:1306.6353 [physics.acc-ph].
22. C. Adolphsen *et al.*, "The International Linear Collider Technical Design Report - Volume 3.II: Accelerator Baseline Design," arXiv:1306.6328 [physics.acc-ph].
23. H. Baer *et al.*, "The International Linear Collider Technical Design Report - Volume 2: Physics," arXiv:1306.6352 [hep-ph].
24. L. Evans *et al.* [Linear Collider Collaboration], "The International Linear Collider Machine Staging Report 2017," arXiv:1711.00568 [physics.acc-ph].
25. K. Fujii *et al.*, "Physics Case for the International Linear Collider," arXiv:1506.05992 [hep-ex].

26. K. Fujii *et al.*, "Physics Case for the 250 GeV Stage of the International Linear Collider," arXiv:1710.07621 [hep-ex].
27. K. Seidel, F. Simon, M. Tesar and S. Poss, "Top quark mass measurements at and above threshold at CLIC," Eur. Phys. J. C **73** (2013) no.8, 2530 https://doi.org/10.1140/epjc/s10052-013-2530-7 [arXiv:1303.3758 [hep-ex]].
28. J. C. Brient and H. Videau, "The Calorimetry at the future e+ e- linear collider," eConf C **010630** (2001) E3047 [hep-ex/0202004].
29. V. L. Morgunov, "Energy flow method for multi - jet effective mass reconstruction in the highly granulated TESLA calorimeter," eConf C **010630** (2001) E3041.
30. J. Brau *et al.*, "International Linear Collider reference design report. 1: Executive summary. 2: Physics at the ILC. 3: Accelerator. 4: Detectors," https://doi.org/10.2172/929487
31. M. A. Thomson, "Particle Flow Calorimetry and the PandoraPFA Algorithm," Nucl. Instrum. Meth. A **611** (2009) 25 [arXiv:0907.3577 [physics.ins-det]].
32. The ILD concept group, see *http://www.ilcild.org*.
33. The SiD concept group, see *http://silicondetector.org*.
34. The 4th detector concept at the ILC, see *http://www.4thconcept.org*.
35. CLIC Detector and Physics Study, see *http://clicdp.web.cern.ch*.
36. CEPC Physics and Detector Working Group, see *http://cepc.ihep.ac.cn*.
37. T. Behnke *et al.*, "The International Linear Collider Technical Design Report - Volume 4: Detectors," arXiv:1306.6329 [physics.ins-det].
38. Y. Banda *et al.*, "Design and performance of improved column parallel CCD, CPC2," Nucl. Instrum. Meth. A **621** (2010) 192. https://doi.org/10.1016/j.nima.2010.05.055
39. Y. Sugimoto *et al.*, "CCD-based vertex detector for GLC," Nucl. Instrum. Meth. A **549** (2005) 87. https://doi.org/10.1016/j.nima.2005.04.032
40. G. Deptuch *et al.*, "Monolithic Active Pixel Sensors adapted to future vertex detector requirements," Nucl. Instrum. Meth. A **535** (2004) 366. https://doi.org/10.1016/j.nima.2004.07.152
41. C. Hu-Guo *et al.*, "First reticule size MAPS with digital output and integrated zero suppression for the EUDET-JRA1 beam telescope," Nucl. Instrum. Meth. A **623** (2010) 480. https://doi.org/10.1016/j.nima.2010.03.043
42. EUDET - Detector R&D Towards the International Linear Collider, see *https://www.eudet.org*.
43. AIDA2020 - Advanced European Infrastructure for Detectors at Accelerators, see *https://aida2020.web.cern.ch*.
44. G. Aglieri Rinella [ALICE Collaboration], "The ALPIDE pixel sensor chip for the upgrade of the ALICE Inner Tracking System," Nucl. Instrum. Meth. A **845** (2017) 583. https://doi.org/10.1016/j.nima.2016.05.016
45. M. Winter, "CMOS Pixel Sensors for ILC related Vertexing and Tracking Devices", presented at American Linear Collider Workshop 2017, SLAC, 2017, https://portal.slac.stanford.edu/sites/conf_public/AWLC17/Pages/default.aspx
46. DEPFET Collaboration for Vertex Detectors at ILC and Belle-II, see *https://www.depfet.org*.
47. A. Nomerotski *et al.* [PLUME Collaboration], "PLUME collaboration: Ultra-light ladders for linear collider vertex detector," Nucl. Instrum. Meth. A **650** (2011) 208. https://doi.org/10.1016/j.nima.2010.12.083
48. F. Luetticke [DEPFET Collaboration], "The ultralight DEPFET pixel detector of the Belle II experiment," Nucl. Instrum. Meth. A **845** (2017) 118. https://doi.org/10.1016/j.nima.2016.06.114
49. I. M. Gregor, "Summary of One Year Operation of the EUDET CMOS Pixel Telescope," arXiv:0901.0616 [physics.ins-det].
50. C. J. S. Damerell and D. J. Jackson, "Design of a vertex detector and topological vertex reconstruction at the future linear collider," *Prepared for 3rd Workshop on Physics and Experiments with e+ e- Linear Colliders (LCWS 95)*, Iwate, Japan, 1995.
51. D. Bailey *et al.* [LCFI Collaboration], Nucl. Instrum. Meth. A **610** (2009) 573 https://doi.org/10.1016/j.nima.2009.08.059 [arXiv:0908.3019 [physics.ins-det]].
52. LCTPC Collabortation, see *https://www.lctpc.org*.

53. K. Aamodt *et al.* [ALICE Collaboration], "The ALICE experiment at the CERN LHC," JINST **3** (2008) S08002. https://doi.org/10.1088/1748-0221/3/08/S08002
54. P. Schade *et al.* [LCTPC Collaboration], Nucl. Instrum. Meth. A **628** (2011) 128. https://doi.org/10.1016/j.nima.2010.06.300
55. R. Bouclier *et al.*, "The Gas electron multiplier (GEM)," IEEE Trans. Nucl. Sci. **44** (1997) 646 [ICFA Instrum. Bull. **1996** (1996) F53]. https://doi.org/10.1109/23.603726
56. M. Killenberg, S. Lotze, A. Munnich, S. Roth, M. Weber and J. Mnich, "Development of a GEM-based high resolution TPC for the International Linear Collider," Nucl. Instrum. Meth. A **573** (2007) 183. https://doi.org/10.1142/9789812773678_0183,10.1016/j.nima.2006.10.396
57. Y. Giomataris, P. Rebourgeard, J. P. Robert and G. Charpak, Nucl. Instrum. Meth. A **376** (1996) 29. https://doi.org/10.1016/0168-9002(96)00175-1
58. D. S. Bhattacharya *et al.*, "A Micromegas-based TPC for the International Linear Collider," DAE Symp. Nucl. Phys. **61** (2016) 962.
59. D. Attie [LC-TPC Collaboration], "Beam tests of Micromegas LC-TPC large prototype," JINST **6** (2011) C01007. https://doi.org/10.1088/1748-0221/6/01/C01007
60. M. Killenberg *et al.*, "Modelling and measurement of charge transfer in multiple GEM structures," Nucl. Instrum. Meth. A **498**, 369 (2003) [arXiv:physics/0212005].
61. L. Hallermann, "Analysis of GEM properties and development of a GEM support structure for the ILD Time Projection Chamber," https://doi.org/10.3204/DESY-THESIS-2010-015
62. M. Lupberger, "The Pixel-TPC: A feasibility study," Thesis, Bonn, 2015.
63. X. Llopart et el., "MediPix2, a 64k Pixel read-out with 55 μm square elements working in single photon counting mode", IEEE Trans. Nucl. Sci. NS-49 (2002) 2279
64. X. Llopart, R. Ballabriga, M. Campbell, L. Tlustos and W. Wong, Nucl. Instrum. Meth. A **581** (2007) 485 Erratum: [Nucl. Instrum. Meth. A **585** (2008) 106]. https://doi.org/10.1016/j.nima.2007.08.079, https://doi.org/10.1016/j.nima.2007.11.003
65. M. Hauschild, "Particle ID with dE/dx at the TESLA-TPC," *Prepared for 5th International Linear Collider Workshop (LCWS 2000), Fermilab, Batavia, Illinois, 24–28 Oct 2000*
66. CALICE Collaboration, see *https://twiki.cern.ch/twiki/bin/view/CALICE/WebHome*
67. J. E. Brau *et al.*, "An electromagnetic calorimeter for the silicon detector concept," Pramana **69** (2007) 1025. https://doi.org/10.1007/s12043-007-0222-2
68. J. Brau *et al.*, "A silicon-tungsten electromagnetic calorimeter with integrated electronics for the International Linear Collider," J. Phys. Conf. Ser. **404** (2012) 012067. https://doi.org/10.1088/1742-6596/404/1/012067
69. G. Nooren *et al.*, "The FoCal prototype - an extremely fine-grained electromagnetic calorimeter using CMOS pixel sensors," JINST **13** (2018) no.01, P01014 https://doi.org/10.1088/1748-0221/13/01/P01014 [arXiv:1708.05164 [physics.ins-det]].
70. F. Sefkow, A. White, K. Kawagoe, R. Pöschl and J. Repond, "Experimental Tests of Particle Flow Calorimetry," Rev. Mod. Phys. **88** (2016) 015003 https://doi.org/10.1103/RevModPhys.88.015003 [arXiv:1507.05893 [physics.ins-det]].
71. F. Sefkow, "The new scintillator-SiPM based analogue HCAL prototype", presented at International Workshop on Future Linear Colliders LCWS2017, Strasbourg, 2017. https://agenda.linearcollider.org/event/7645/overview.
72. C. Adloff, J. Blaha, J. J. Blaising, M. Chefdeville, A. Espargiliere and Y. Karyotakis, "Monte carlo study of the physics performance of a digital hadronic calorimeter," JINST **4** (2009) P11009 [arXiv:0910.2636 [physics.ins-det]].
73. V. Buridon *et al.* [CALICE Collaboration], "First results of the CALICE SDHCAL technological prototype," JINST **11** (2016) no.04, P04001 https://doi.org/10.1088/1748-0221/11/04/P04001 [arXiv:1602.02276 [physics.ins-det]].
74. S. George [ATLAS Collaboration], "Design and expected performance of the ATLAS trigger and event selection," Eur. Phys. J. direct **4** (2002) no.S1, 06. https://doi.org/10.1007/s1010502cs106
75. S. Dasu [CMS Collaboration], "CMS trigger and event selection," Eur. Phys. J. direct **4** (2002) no.S1, 09. https://doi.org/10.1007/s1010502cs109

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.